

# The Regular Derivation in Serbian Principles and Classification Using NooJ

Miloš Utvić  
Faculty of Philology,  
University of Belgrade  
misko at matf bg ac yu

# Contents

- Unknown word in Serbian
- Regular derivation in Serbian
- Implementing regular derivation in e-dictionaries of Serbian (NooJ, Prolex)
- Concept of superlemma
- Classification of regular derivational paradigms of toponyms

# Unknown word

- Words not present in a electronic dictionary but found in unrestricted texts during a morphological analysis.
- Types of unknown words in Serbian:
  - text-specific words (proper names representing fictional characters, sequences of foreign language words ...),
  - missing words (name entities, abbreviations, dialect words ...)
  - results of regular derivation (gender motion, )

# Regular derivation

- Class of derivational processes which induce change to the lexical meaning in a predictable way

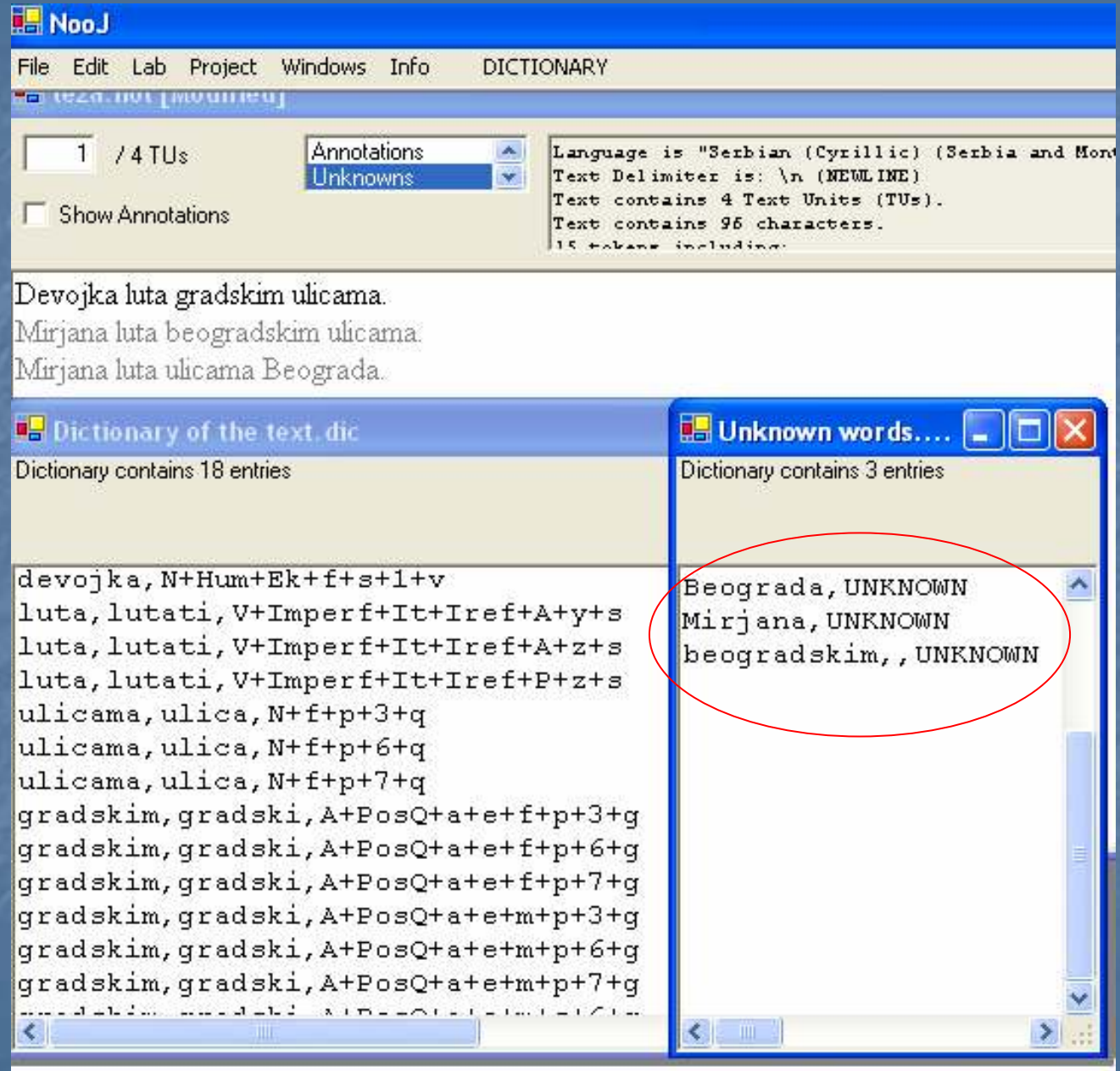
gender motion	amplification of meaning (diminutives, augmentatives)	poss. and relational adjectives	verbal nouns
<i>Nišlija</i> > <i>Nišlijka</i>	<i>kuća</i> > <i>kućica</i> (dim.) > <i>kućetina</i> (aug.)	<i>Milan</i> > <i>Milanov</i> <i>Nada</i> > <i>Nadin</i> <i>Niš</i> > <i>niški</i>	<i>pričati</i> > <i>pričanje</i>

# Reg. derivation in morphological e-dictionary of Serbian

- Results of reg. derivation represent a broad category of unknown words in Serbian (including results of regular derivation from proper names)
- Systematic incorporation of regularly derived lemmas into the e-dictionary:
  - multiplies the size of e-dictionary
  - complicates its maintenance
  - adds considerably to the text ambiguity
  - loses relations between basic word and its derivatives, so dictionary can't be used for the analysis of synonymy relations
- Incorporation of only those regularly derived lemmas which are present in paper dictionaries leads to serious inconsistencies

# Example

- Devojka luta gradskim ulicama.  
(Girl wanders city streets.)
- Mirjana luta beogradskim ulicama.  
(Mirjana wanders Belgrade streets.)
- Mirjana luta ulicama Beograda.  
(Mirjana wanders streets of Belgrade).



## Example 2

- General Secretary of the Communist Party of France Robert Hue ...  
(*Generalni sekretari Komunističke partije Francuske Rober I ...*)
- surname *I (Hue)*
- roman number ("the first")
- conjunction "and"

# Prolex

- Since 1996, the Prolex project concerns proper names processing, particularly toponyms and inhabitant names, and stresses the need to link proper names together.
- Today, the main motivation of the Prolex project is to develop a multilingual dictionary of proper names and their relationships.
- Resources of proper names are developed for several European languages, including Serbian

# Prolex

```
<struct type="Prolex">  
  <struct type="pivot">  
    <feat type="identifier">48715</feat>  
    <struct type="prolexeme">  
      <feat type="language">sr</feat>  
      <feat type="lemma">  
        Beograd  
      </feat>  
      <feat type="pos">name</feat>  
      <feat type="category">proper name</feat>  
    </struct>  
  </struct>  
</struct>
```

# Prolex levels (layers)

## Paris

Type : Ville  
Existence : Historique  
Nom relationnel : Parisien  
Adj. Relationnel : parisien  
Holonyme : Île-de-France  
Accessibilité : France (Capitale), Île-de-France (Capitale)  
Classifiant : chef-lieu, préfecture de région, ville  
Traduction : Paris (allemand), Parigi (italien), Paris (portugais),  
París (espagnol), Parijs (hollandais), Paris (anglais),  
Pariz (serbe), ?? (coréen)  
Encyclopédie : Wikipédia  
Plus de détails

## Pâris

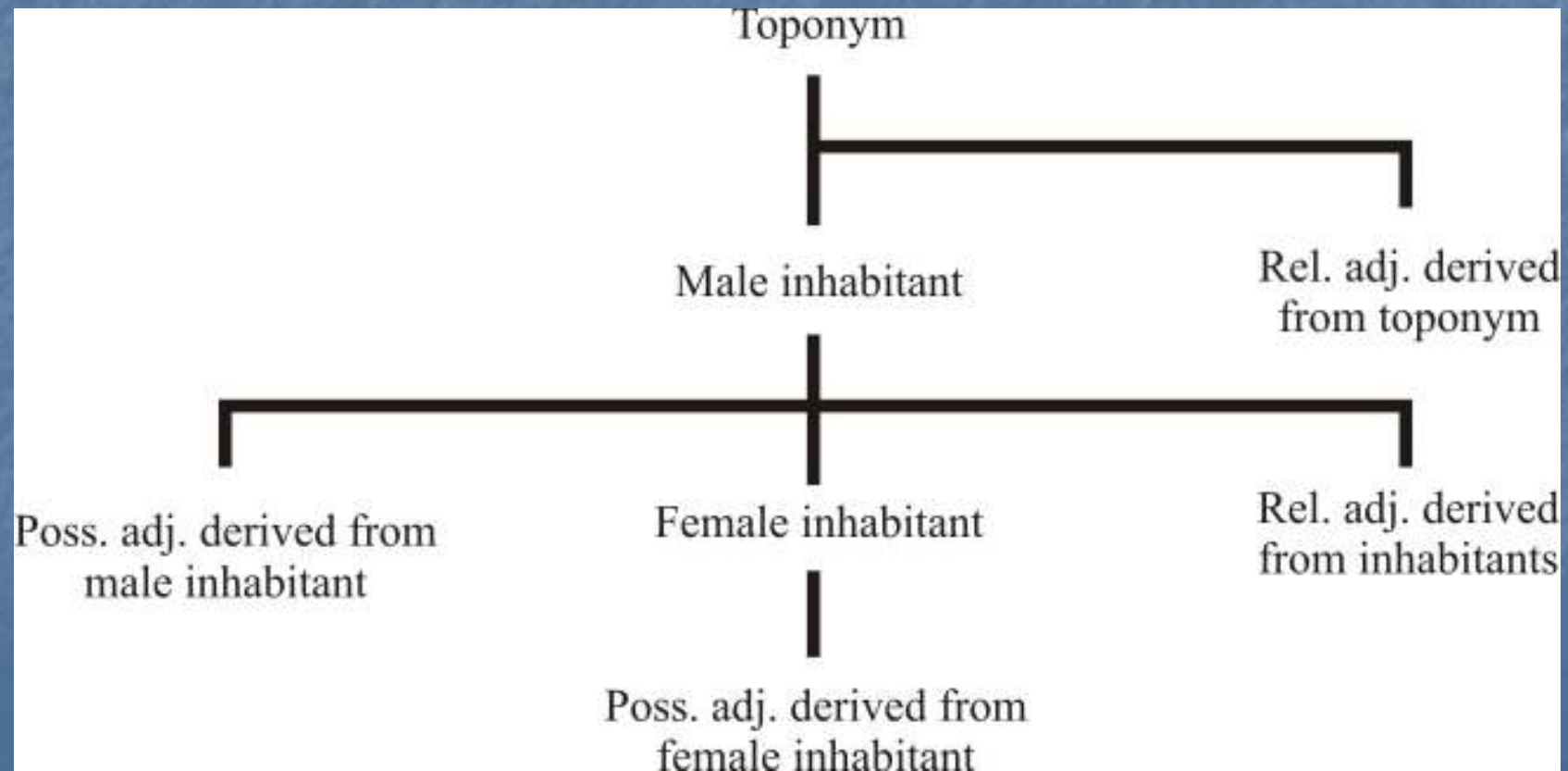
Type : Célébrité  
Existence : Historique  
Holonyme : Grèce antique  
Encyclopédie : Wikipédia  
Plus de détails

Résultat contenant la séquence

[Aéroports de Paris](#)  
[Commune de Paris](#)

# General derivational hierarchy (regular derivation from toponyms)

- Inflection lemma and “derivational” lemmas



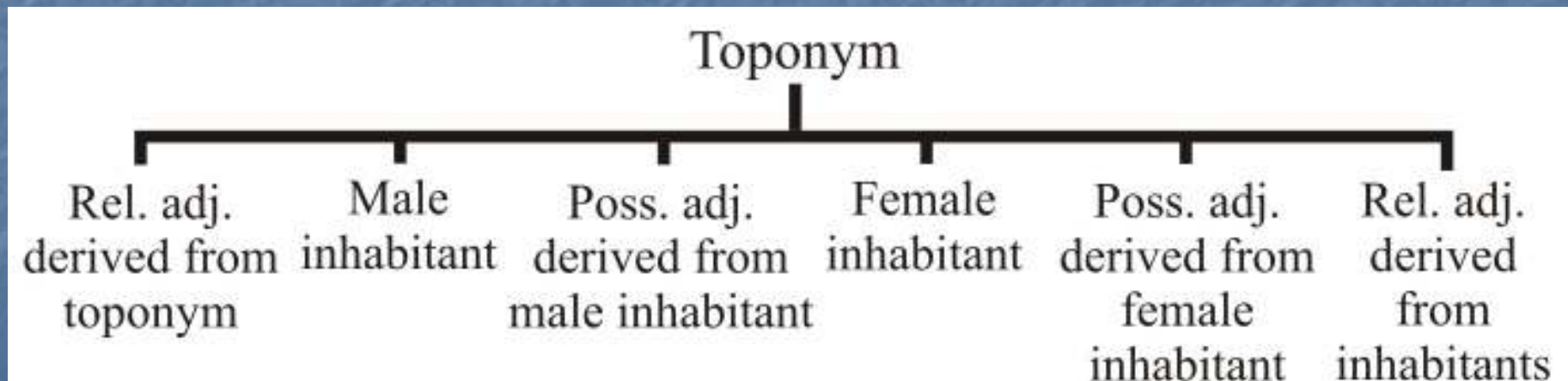
# Example of derivational hierarchy for toponym *Pariz* (*Paris*)



- (*turski* > *Turska* (Turkish > Turkey), *grčki* > *Grčka*)

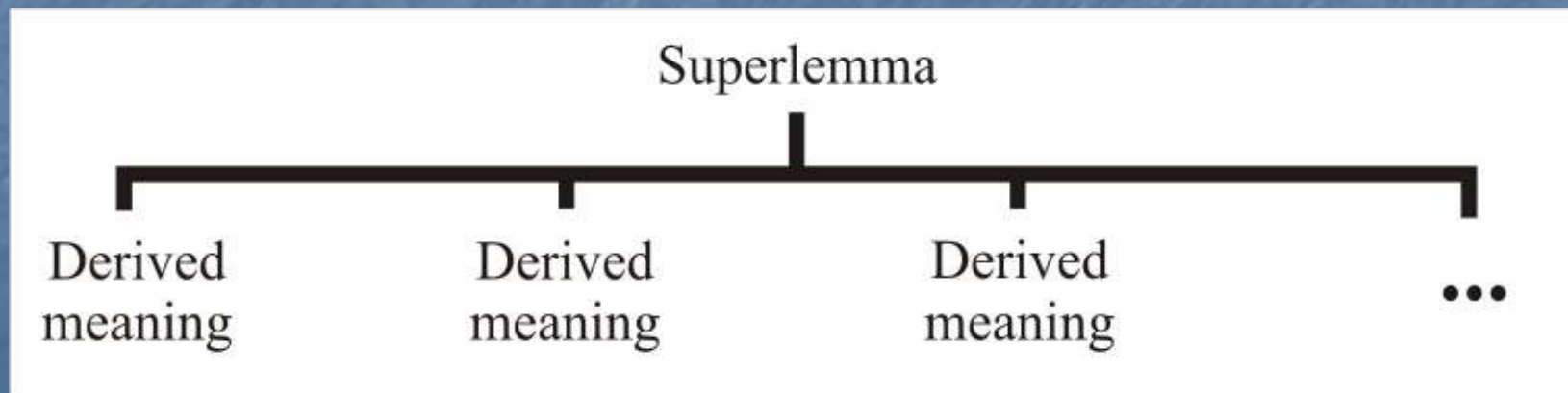
# Hierarchy of meanings

- Hierarchy of meanings instead of hierarchy of derived forms (eg. "toponym X", "which relates to X", "male inhabitant of X", "which belongs to male inhabitant of X", "which relates to all inhabitants of X" etc.)



# Superlemma

- Superlemma = “basic meaning” from which all other meanings are derived.
- The order in which derivations happen isn't important, only derived meanings are relevant (these meanings are predictable in case of the regular derivation).



# Derivational suffixes (toponyms)

Derived forms	Derivational suffixes	Inflection class
Rel. adjectives	<i>-ski, -ški, -čki, -ćki</i>	A2
Poss. adjectives	<i>-ov, -ev, -in</i>	A1
Female inhabitant	<i>-ka, -inja, -ica</i>	N661 N601 N651
Male inhabitant	<i>-ac, -in (-anin, -janin), -ar, -ak, -lija, -∅, ...</i>	N42, N60, N2, N10, N741,...

# Principles of classification

- How to describe derivational paradigm?
- What are the “correct” names of male and female inhabitants and related adjectives
  - paper dictionaries and orthography;
  - local names (how inhabitants call themselves, *Puležani* and *Puljani*);
  - newspapers
    - *Tuzlak, Tuzlanin, Tuzlanac*
    - Dilemma: *-ac* or *(j)anin* (*Jamajkanac* or *Jamajčanin, jamajkanski* or *jamajčanski*)
    - *Somalac/Somalijac, Bask/\*Baskijac*

# Dublets

- Sometimes there are pairs of adjectives, one motivated by toponym (*Beograd* > *beogradski*) and the other one motivated by inhabitants (*Beograđani* > *beograđanski*)
- Paper dictionaries are inconsistent (RMSMH i RSANU)
  - *banatski/banaćanski* (different meanings)
  - *norveški/norvežanski* (the same meanings)
  - *meksički/meksikanski* (the first relates only to *Mexico*, while the second relates both to Mexico and Mexicans)
- *portugalski* | ∅ or *portugalski* / *portugalski*  
∅ | *vojvođanski* or *vojvođanski* | *vojvođanski*

# Phonetic alternations

- produce more sophisticated differentiation of toponyms and allomorphs of suffixes  
(e.g. *-ski*, *-ški*, *-čki*, *-čki*)
- Jotation (*Banat* > *Banaćanin*,  $t+j=ć$   
*Tajland* > *Tajlandžanin*,  $d+j=đ$ )
- Palatalization (*Lika* > *Ličanin*)
- Voicing and devoicing (*Šabac* > *Šapčanin*)
- Consonant loss or elision  
(*Perast* > *peraški*)
- Operators which simulate phonetic alternations in order to decrease the number of classes
  - automatic jotation  
<J>:  $t \Rightarrow ć$   
<J>:  $d \Rightarrow đ$
  - automatic voicing and devoicing  
*Sabac* > *Šapčanin*  
Leskovac > Leskovčanin

# Sources used for description of derivational paradigms of toponyms

Toponyms				Derivational paradigms			
All resources (e-dictionaries and the rest)		e-dictionaries of toponyms		Corpus of contemporary Serbian			
SWU	MWU	SWU	MWU	SWU		MWU	
				<i>-ac</i>	<i>-anin</i>	<i>-ac</i>	<i>-anin</i>
12788	2523	992	215	293	331	104	80
				624		184	
15311		1207		806			

# NooJ dictionaries of toponyms

lemma,PoS+FLX=Cxx{+DRV=Dxx[:Fxx]}{+SynSem}

London,N+FLX=N1001+NProp+Top+IsoUKgr

# Derivation in NooJ dictionaries

- Crna\_Gora,N+FLX=CGFlx+DRV=CGDrv  
+NProp+Top
- CGDrv = <P><B>o<S><R><C><RW><B>  
(ac/N:AC + cyev/A:EV + ka/N:KA + kin/A:IN)  
+<P><LW><R><C><RW><B>o<S><R><C><RW>  
<B>ski/A:SKI;

# NooJ textual rewriting rules describing derivational paradigm

- <P><B>o<S><R><C><RW><B>ac
- Crna Gora\_
- Crna\_Gora (after applying the operator <P>)
- Crn\_Gora (after applying the operator <B>)
- Crno\_Gora (after insertion of the connect. vowel *o*)
- CrnoGora (after applying the operator <S>)
- CrnoGora (after applying the operator <R>)
- Crnogora (after applying the operator <C>)
- Crnogora\_ (after applying the operator <RW>)
- Crnogor\_ (after applying the operator <B>)
- Crnogora\_ (after insertion of the character *a*)
- Crnogorac\_ (after insertion of the character *c*)

# Suggestions for the improvement of derivation model

- *Crnogorca, Crna\_Gora, N+Inh+Hum+FLX=CGFlx*  
*+DRV=CGDrv*  
*+NProp+Top*  
*+m+s+2*

*crnogorskog, Crna Gora, A+FLX=CGFlx*  
*+DRV=CGDrv*  
*+NProp+Top*  
*+m+s+2*

- Insufficient readability of generated forms:
  - Information about derived lemmas is lost (*Crnogorac, crnogorski*)
  - Mix of semantic properties relating only to **superlemma** and those relating only to **derived forms**
  - XML Format of dictionary?

# Classification

- For each superlemma and its derivational paradigm program geord automatically constructs corresponding NooJ textual rewriting rule. That rule describes necessary transformations of toponym lemma which generate its derivational paradigm.
- All toponyms sharing the same rule are elements of one derivational class described by that rule.

# Rule for SW superlemma (toponym *Austrija, Austria*)

- Rule:

<B>anac/**N:AC** + <B>ančev/**A:EV** +  
<B>anka/**N:KA** +  
<B>kin/**A:IN** +  
<LW><R><C><RW><B>ski/**A:SKI**;

# MWU (2-WU) Toponyms and simple derived forms

- Types:
  - (type 1) the first word unit doesn't affect derivation (*Herceg **Novi*** > ***novljanski***);
  - (type 2) the second word unit doesn't affect derivation (***Homoljske** planine* > ***homoljski***);
  - (type 3) both word units affect derivation (***Crna Gora*** > ***crnogorski***). Derived forms are 1-WU compounds which often have a vowel ('o' ili 'e') connecting the parts of superlemma word units.

# MWU -> SWU derivation rule

- *For the sake of simplicity POS and inflection codes are omitted*
- *Crna Gora >*  
*crnogorski + Crnogorac + Crnogorčev*  
*+ Crnogorka + Crnogorkin*

<P><LW><R><C><RW><B>o<S><R><C><RW><B>ski  
+ <P><B>o<S><R><C><RW><B>(ac + čev + ka + kin)

# Classification results (simple words)

jednočlani toponimi	ukupan broj toponima	broj klasa	broj singltona	broj toponima u 5 najbrojnijih klasa				
				93	79	27	21	7
etnik na <i>-ac</i>	293	46	31	93	79	27	21	7
etnik na <i>-in</i> ( <i>-(j)amin</i> )	331	81	41	40	34	14	13	13
ostali etnici	34	26	18	2	2	2	2	2

# Classification results (MWU)

dvočlani toponimi	ukupan broj toponima	broj klasa	broj singltona	broj toponima u 5 najbrojnijih klasa				
etnik na <i>-ac</i>	104	25	11	24	12	12	10	7
etnik na <i>-in</i> ( <i>-(j)anin</i> )	80	43	26	7	5	5	4	4
ostali etnici	2	2	2	1	1			

# Conclusion

- This approach enables more precise and systematic description of regular derivation in e-dictionaries of proper names in Serbian. Still, there are a few problems which wait the solution.
- Goal: description of regular derivation classes in Serbian in general (not only for proper names) in a way which is independent of any implementation (Prolex, NooJ etc.)

Thank you!

