SW4CH 2017
Nicosia, Cyprus, September 24-27, 2017

Using a Wordnet Ontology to Improve the Search of the Digital Dialect Dictionary

Miljana Mladenović, eVox Solutions, Belgrade, Serbia
Ranka Stanković, University of Belgrade, Faculty of Mining and Geology
Cvetana Krstev, University of Belgrade, Faculty of Philology

We will present…

- Method for automatic relating between dialect term and corresponding terms in standard language, www.vranje.co.rs
- The method uses SWRL rules defined in the Serbian WordNet ontology to identify sets of synonymous words.
- It also uses e-dictionaries to produce correct lemmas in the standard language that users usually use for search.
- The method was applied and evaluated on verbs and a group of nouns derived from verbs (verbal nouns).
- We compared results obtained by the system with human evaluators and achieved the accuracy of 89.7%.

Digital dictionary of the South Serbian dialect

http://www.vranje.co.rs

1st

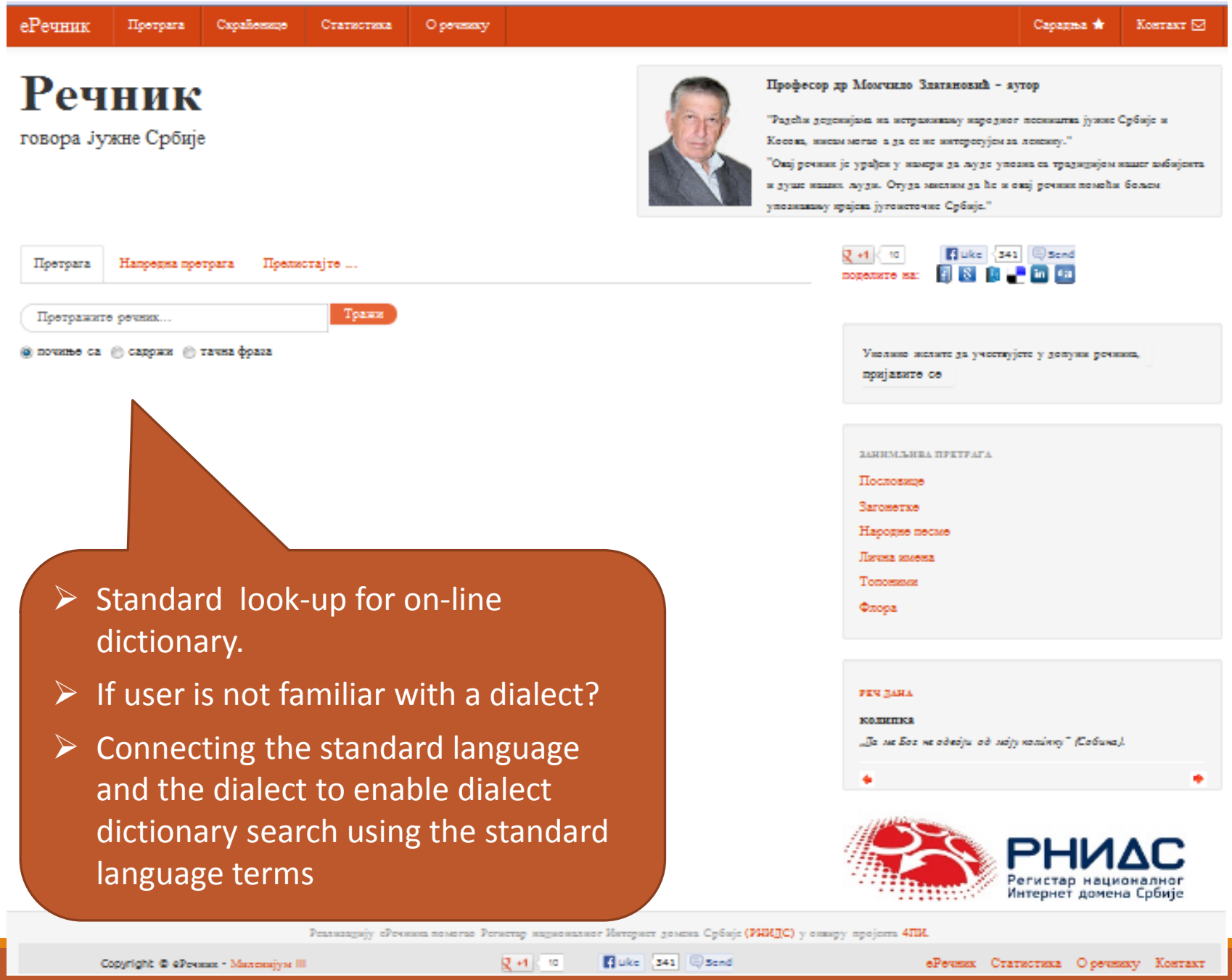
- implementation of an on-line dialect vocabulary for Serbian, produced from traditional dialect dictionaries

~20,000 entries:

- POS, linguistic information, sound (pronunciation), usage examples, dialect phrases, geolocation, etymology, semantic data, social networks and crowdsourcing.

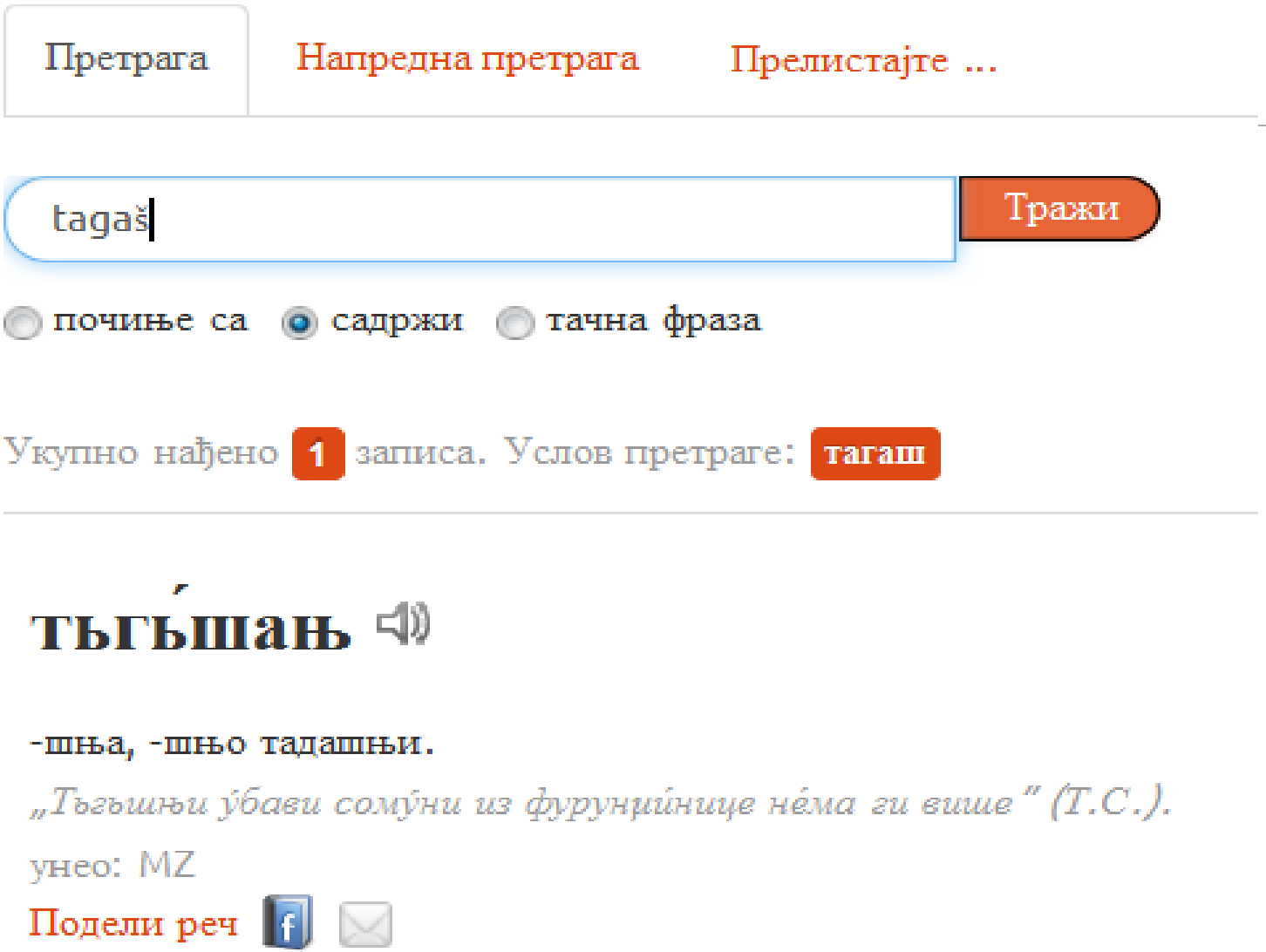
Search

- by a term, by boolean metadata queries
- browsing by the 1st letter

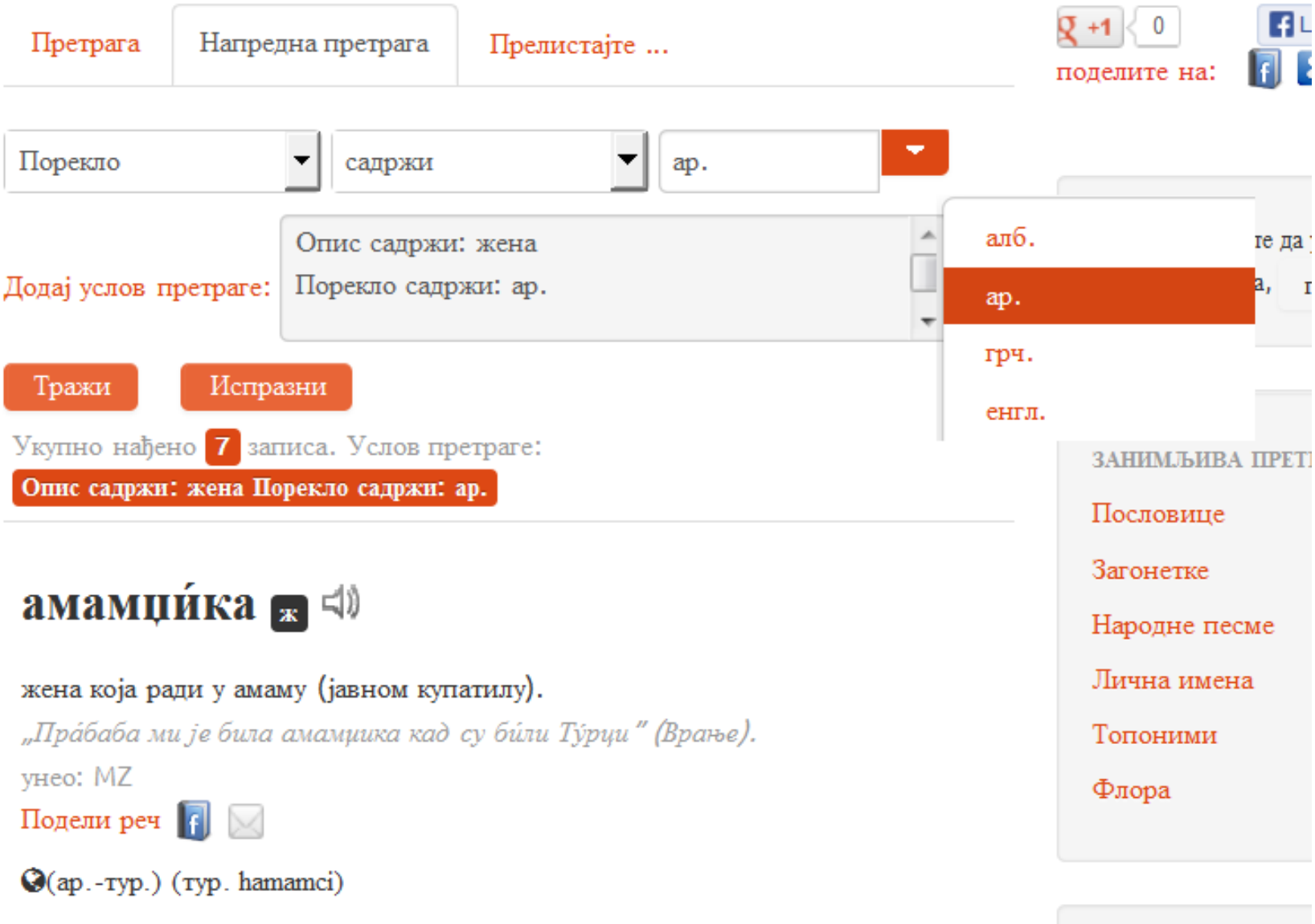


- Standard look-up for on-line dictionary.
- If user is not familiar with a dialect?
- Connecting the standard language and the dialect to enable dialect dictionary search using the standard language terms

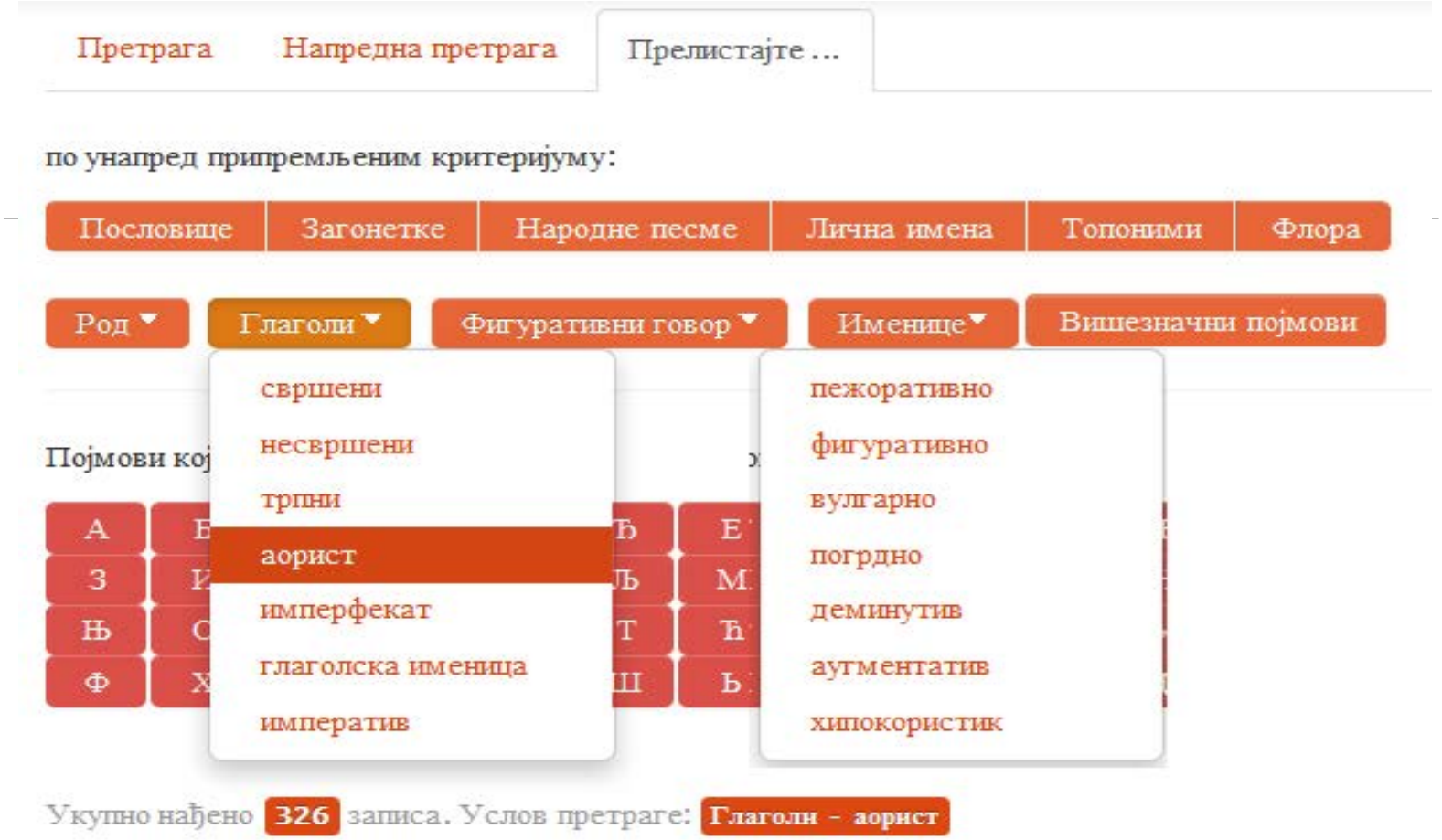
Typical keyword based search



Boolean query…



Semantic search



First letter search (filter)

Претрага | Напредна претрага | Прелистајте ...

по унапред припремљеним критеријуму:

Пословице | Загонетке | Народне песме | Лична имена | Топоними | Флора

Род | Глаголи | Фигуративни говор | Именице | Вишезначни појмови

Појмови који почињу словом:

А	Б	В	Г	Д	Ђ	Е	Ж
З	И	Ј	К	Л	Љ	М	Н
Њ	О	П	Р	С	Т	Ћ	У
Ф	Х	Ц	Ч	Џ	Ш	Ь	Дз

Укупно нађено 69 записа. Услов претраге: Дз

дзьвни

одјекује.

„Удара сас чук, па све дзьвни" (Владовце). „Петлови појев, мори, Морава дзьвни" (нар. пес.) (Врање)

Geolocated search results

ашл Тражи

○ почиње са ◉ садржи ○ тачна фраза

Укупно нађено 2 записа. Услов претраге: ашлак

ашльк м

мали трошак.

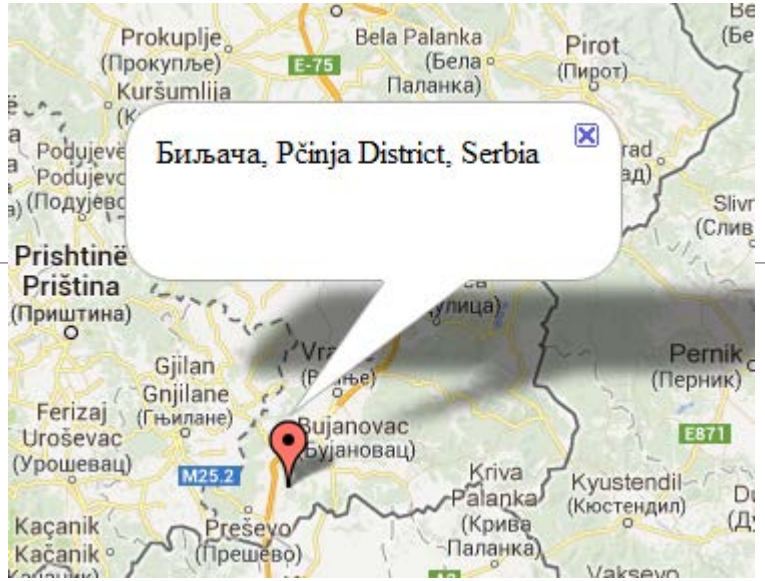
„Дај неки динáр, да ми се нáђе за ашльк" (Биљача).

унео: MZ

Подели реч

(ар.-тур.) (тур. harçlık)

Биљача



цицијашльк м

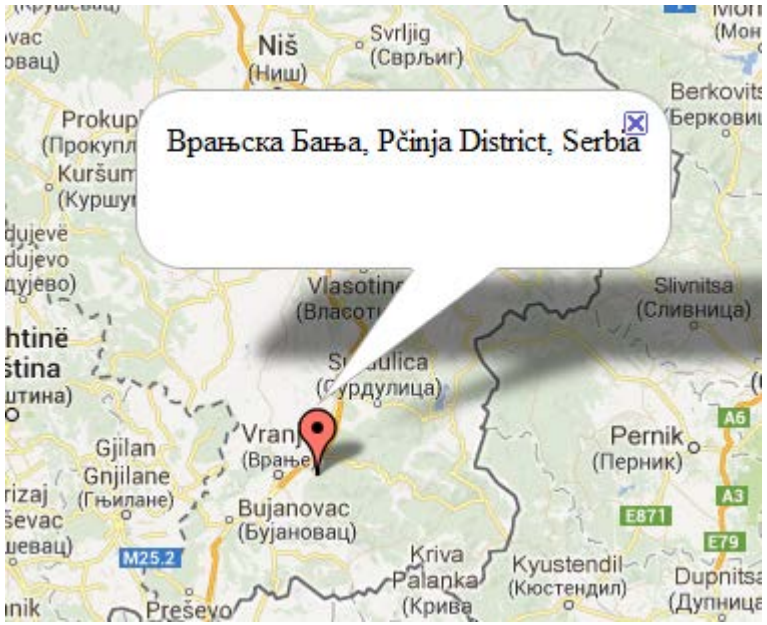
тврдичлук.

„Од њóјан цицијашльк пóголем га нéма" (Врањска Бања).

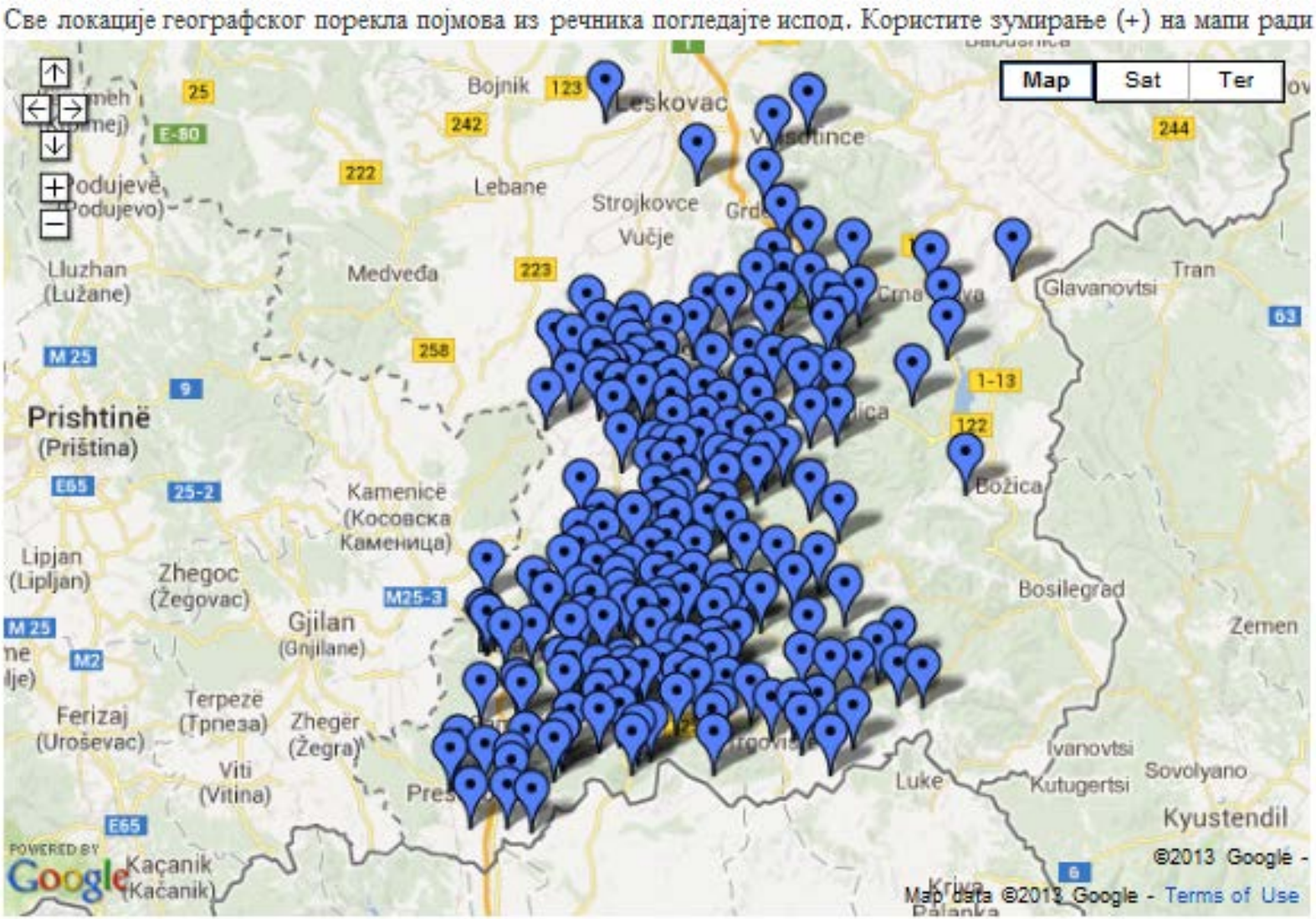
унео: MZ

Подели реч

Врањска Бања



Lexical entry geolocation



Resources for improvement of searching performances

- Serbian morphological e-dictionaries and grammars
  - to produce all inflected forms of standard terms
  - 140,000 lemmas & 5 million forms; 18,000 multi-word lemmas
- Serbian WordNet (SWN) OWL2 ontology
  - rules expressed in Semantic Web Rule Language (SWRL) to generate synonymous groups on the basis of the indirect synonymy relation.
- University of Belgrade
  - Human Language Technology Group

Use of morphological e-dictionaries

- Headword of the verb entry is the present tense, first person singular
- User search for verbs using ifinitive
- Infinitive form (lemma) of dialect verb and verb in the standard Serbian (from definition) was added
- After separation of all synonyms aligned with a dialect, infinitive forms were attached to the original form.
- For 3,452 verb entries 7,353 synonyms were detected

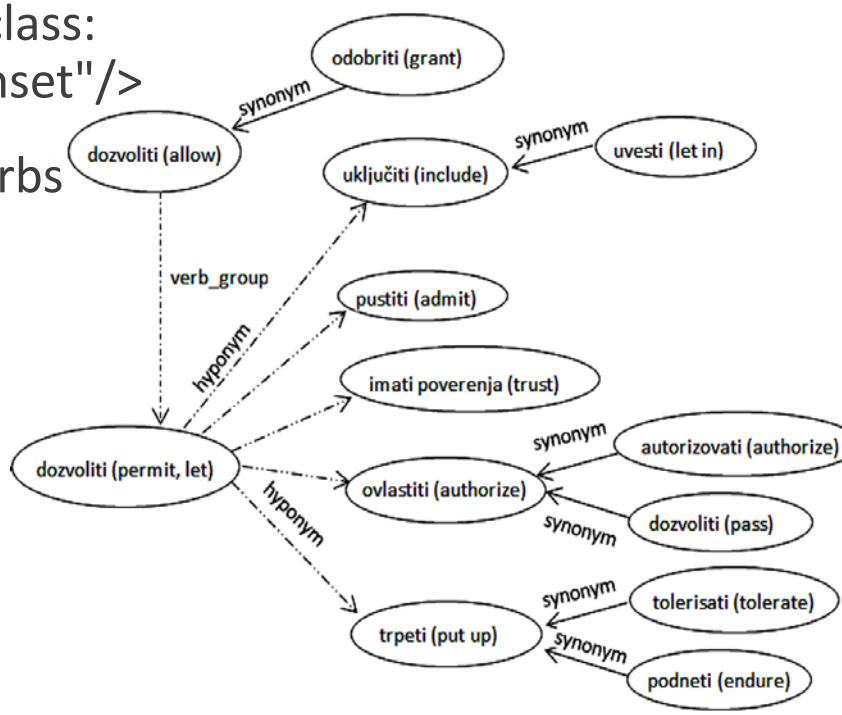
- batalim_bataliti | batalen, ostavim_ostaviti, napustim_napustiti
- batisujem | kvarim_kvariti, upropašćujem_upropašćivati
- bednim se | lepo se odevam_odevati, doterujem_doterivati se
- begam_begati | begaj, ja bega_begati, ti bega_begati, begajeći, bežim_bežati

Use of morphological e-dictionaries

- Lemma was assigned for 505 dialect forms out of 3,452 dialect forms given in first person singular, present tense.
- Infinitive forms were assigned to 4,384 word forms in standard Serbian that were connected to dialect forms (out of 7,353).
- Not lemmatized words that consisted of word not presented in e-dictioanries, or adjectives used to describe verbs
- Relation between verbal nouns and verbs was established in some entries but not systematically.
- In e-dictionaries all verbal nouns are marked with a special marker -> 700 relation were established.

Finding the set of near synonyms by using the WordNet ontology

- Serbian WordNet (SWN), based on Princeton WordNet (PWN) has more than 22,000 concepts (synsets)
- SWN ontology has currently 2,243 verb synsets defined as ontology individuals belonging to the VerbSynset class:
  <rdf:type rdf:resource="&swn30;VerbSynset"/>
- Rules: generate synonymous pairs of verbs found in the SWN ontology not based only on the relation of direct synonymy.
- Broader set of synonyms for each verb defined in SWN ontology produsedusing relations: synonym, similar to, also see, verb group, hyponym.



Reasoning rules in the SWN ontology

- Eclipse Java EE IDE Luna and Apache Jena for reasoning at the level of OWL 2 language by converting OWL rules into the Jena rules format.

"[rule1:(?a eg:label ?b)(?a eg:synonym ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]"

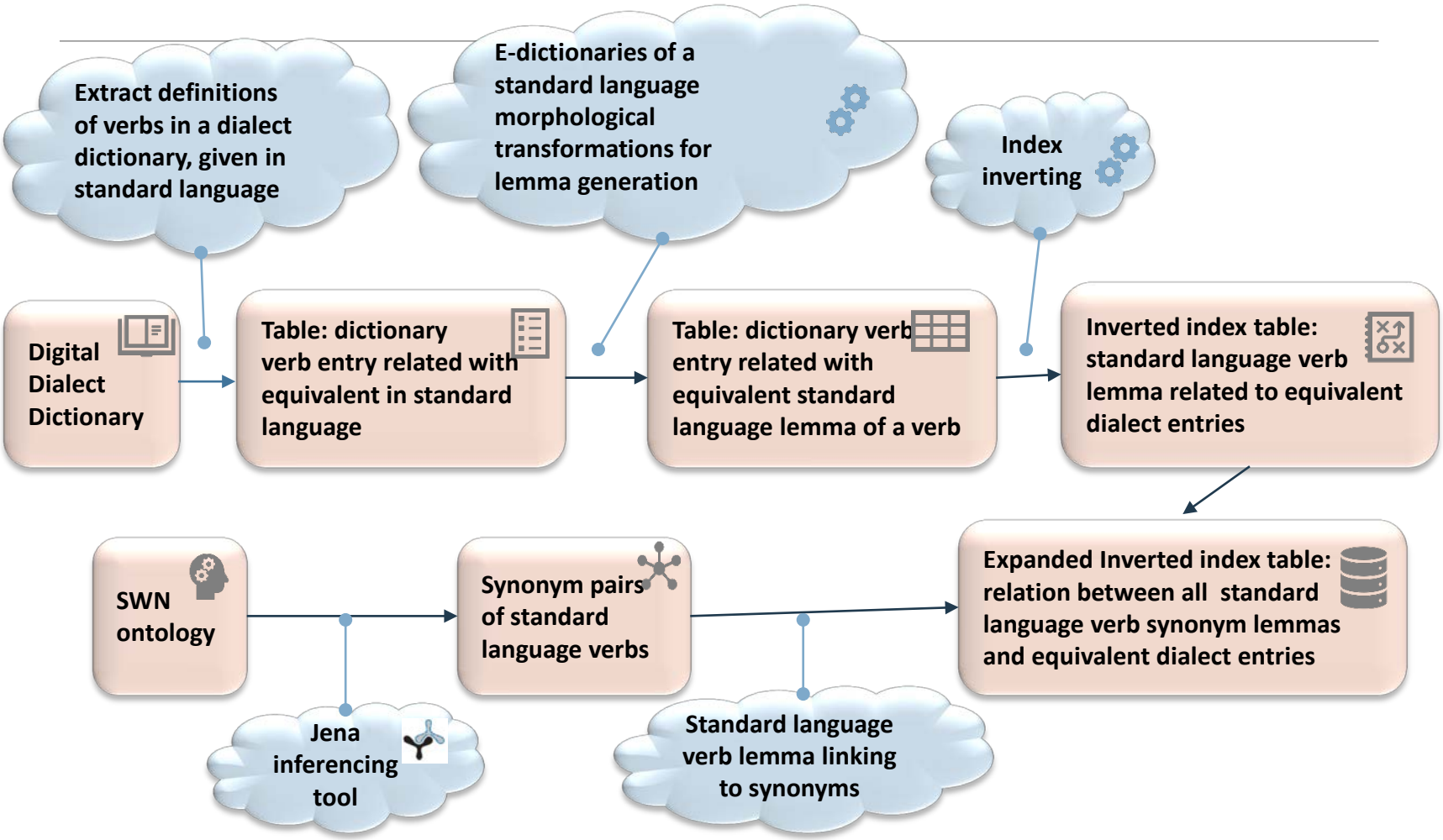
"[rule2:(?a eg:label ?b)(?a eg:similar_to ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]

.....

"[rule6:(?a eg:similar_to ?c)(?a eg:label ?b)(?c eg:synonym ?d)
(?d eg:label ?e) -> (?b eg:indirectSynonymy ?e)]„

- 33 reasoning rules for indirectSynonymy relation
- after inferencing, 6,430 indirectSynonymy related pairs of verbs.

Architecture of the system for building a resource that improves the dialect dictionary search tool



Example

1) Definition extraction

- isabim
  "(imp. isabi; aor. ja isabi, ti isabi; r.pr. isabija, -ila, -ilo) svr. iskvarim, upropastim.„

2) Lemmatization

- isabim |
  isabi; ja isabi; ti isabi; isabija; iskvarim_iskvariti; upropasti_upropastiti

3) Inverted table

- upropastiti |
  isabim batišem dokrajišem istrovim izabim izakam oznobim profućkam

4) Inference rules

- upropastiti |
  unerediti, uništiti, uprskati, zabrljati, zakrmačiti, zasvinjiti

5) Join

- Upropastiti, unerediti, uništiti, uprskati, zabrljati, zakrmašiti, zasvinjiti | isabim, batišem, dokrajišem, istrovim, izabim, izakam, oznobim, profućkam

Evaluation

- Estimation of the accuracy of pairing the DD and SL entries: 2 language experts annotated the inverted (step 3)
  - ✓ Infinitive SL has similar meaning as DD verb ?
    - 1 - yes
    - 2 - not clear
    - 3 - no
- Automatic procedure: DD headwords not related to any infinitive
- Infinitive classified ~ take a part in relations 1) related 2) unrelated
  - Human marks 1 with related $\Rightarrow$ true positives.
  - Human marks 2 and 3 compared to related $\Rightarrow$ false positives.
  - Comparing with the unrelated set $\Rightarrow$ false and true negatives.

Evaluation

	System Yes	System No
Expert yes	$tp = 3022$	$fn = 436$
Expert no	$fp = 0$	$tn = 784$

The confusion matrix

- whether dictionary entries are correctly aligned with standard language entries
  - P = tp=(tp + fp) = 1.000
  - R = tp=(tp + fn)) = 0.874
  - F1 = 2PR=(P + R) = 0.933
  - Accuracy= 0.897

Remarks

- method is completely precise
- FN: shortcomings in the DD
  - typos, non-standard verb forms,
  - missing SL verb in definition,
  - misineterpreted DD verb

Conclusion

Method for improving search of the DD with key-terms in SL

- SL e-dictionaries lemmatize verb forms
- Serbian WordNet based SWRL rules identifies sets of synonymous words for each verb and verbal noun defined in the ontology
- Join two sets of synonym words (from DD and from SL)

Evaluation of the method with data provided by humans

- Accuracy =89.7%.

Future work

- experiment with other POS
- try to expand the set of ontological rules used in this system