



Vebran API for query expansion in Serbian Corpora

Ranka Stanković, Miloš Utvić
University of Belgrade, Serbia

Serbian Uniterx Day

11.3.2019

Introduction



- Goals:
 - To upgrade the existing web search interfaces for language resources;
 - To enable querying language resources supported with available lexical resources.
- Language resources to be searched:
 - Digital libraries and
 - Corpora.
- Lexical resources to support the search:
 - Morphological electronic dictionaries;
 - WordNets;
 - Terminological databases.
- Implementation: Vebran web API

Language resources



Digital libraries

- Biblisha
- Nara
- ROMeka
- ...

Monolingual corpora:

- **SrpKor2013 (122 MW)**
- SrpKor2003 (22 MW)
- SrpLemKor
- **RudKor (2.7 MW)**
- ...

Bilingual and multilingual corpora:

- SrpFranKor (b)
- SrpEngKor (b)
- SELFEH (b)
- Verne80 (m)
- 1984 (m)
- SrpNemKor (b)
- ...

Corpora



Development:
University of Belgrade &
Jerteh

- as a means of improving the search of the digital library based on linguistic annotation, and
- as a resource for various linguistic and terminological research, including extraction.

Three different systems for
diverse types of usage
scenarios

- Unitex, used to create corpus for custom information extraction tasks
- IMS Open Corpus Workbench (CWB) and an adaptation of CQPweb, a web-based graphical user interface and
- NoSketch Engine

Digital libraries



Omeka Classic

- Annales géologiques de la Peninsule balkanique - Geološki anali Balkanskog poluostrva (1889 - 2018) <http://gabp-dl.rgf.rs/> (public)
- Romeka – mining documentation <http://romeka.rgf.rs/> (mostly public, projects private)

OmekaS (semantic)

- Mining IS project documentation <http://eps.rgf.rs> (private)

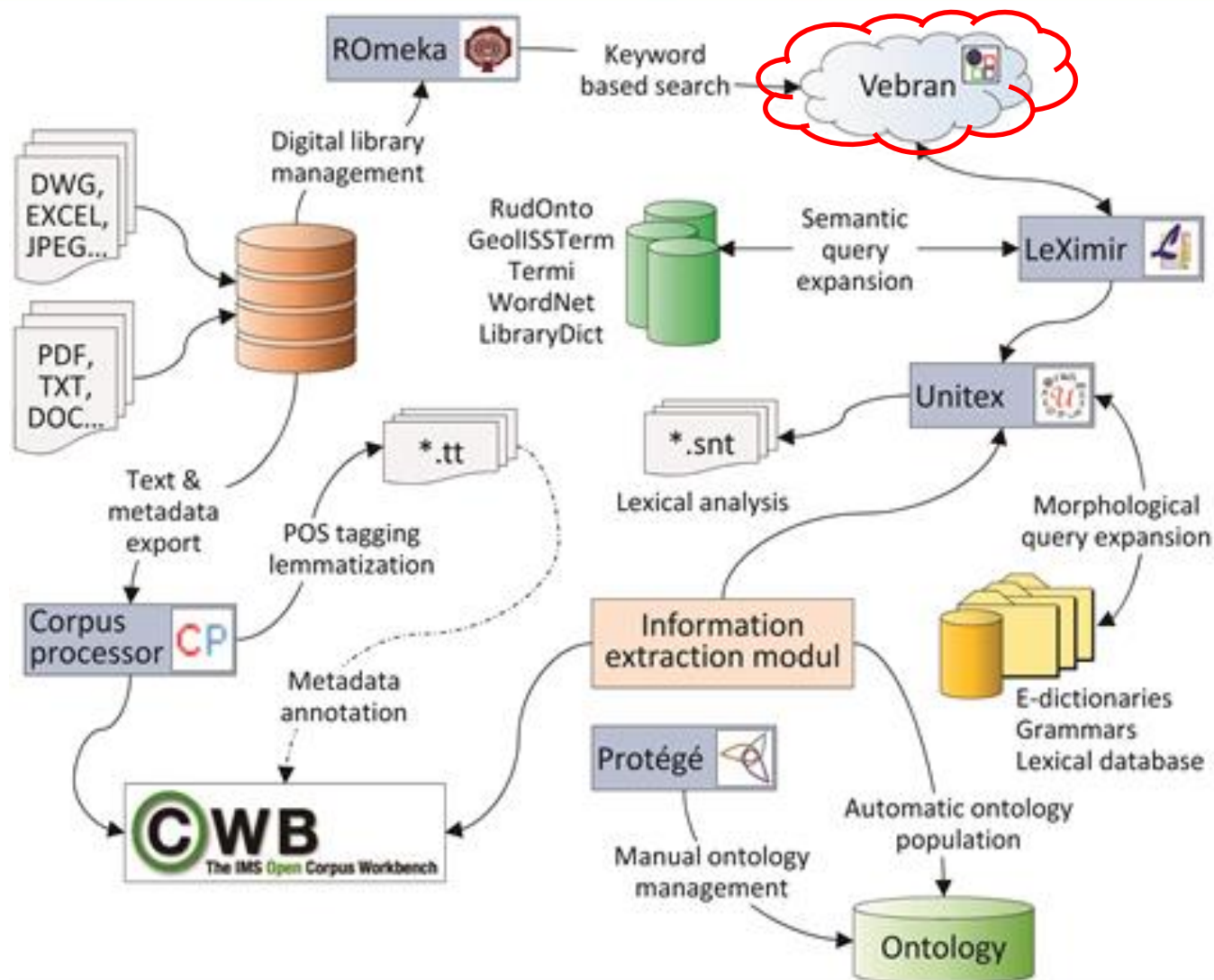
Dspace

- NaRA - National Repository for Agricultural Education <http://arhiva.nara.ac.rs/> (developed and maintained by RCUB & all Serbian faculties of Agriculture)


In development

- RSANU bibliography sources
- Serbian corpus bibliography

Vebran API exchange of data



What are we upgrading?



Existing web
interfaces for
searching
corpora

- IMS CWB
- NoSketch Engine

Serbian corpora

- SrpKor: contemporary Serbian texts
- RudKor: professional texts from the mining area

How do we upgrade?



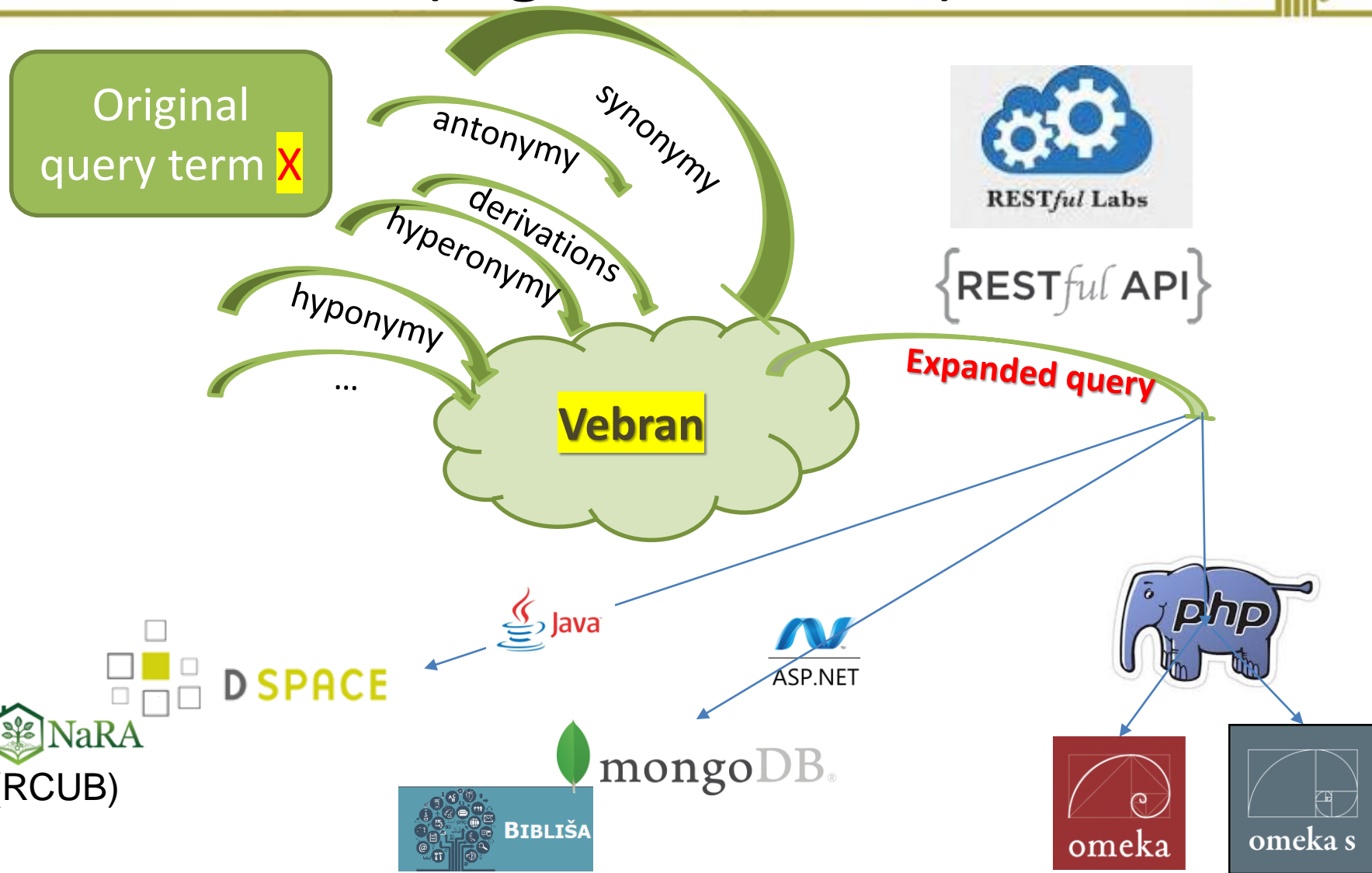
Query expansion based on external lexical resources for Serbian

- Morphological electronic dictionaries (developed for Unitex)
- Semantic network Wordnet (Serbian and English)
- Terminological databases Termi, RudOnto, GeolISS, RBI

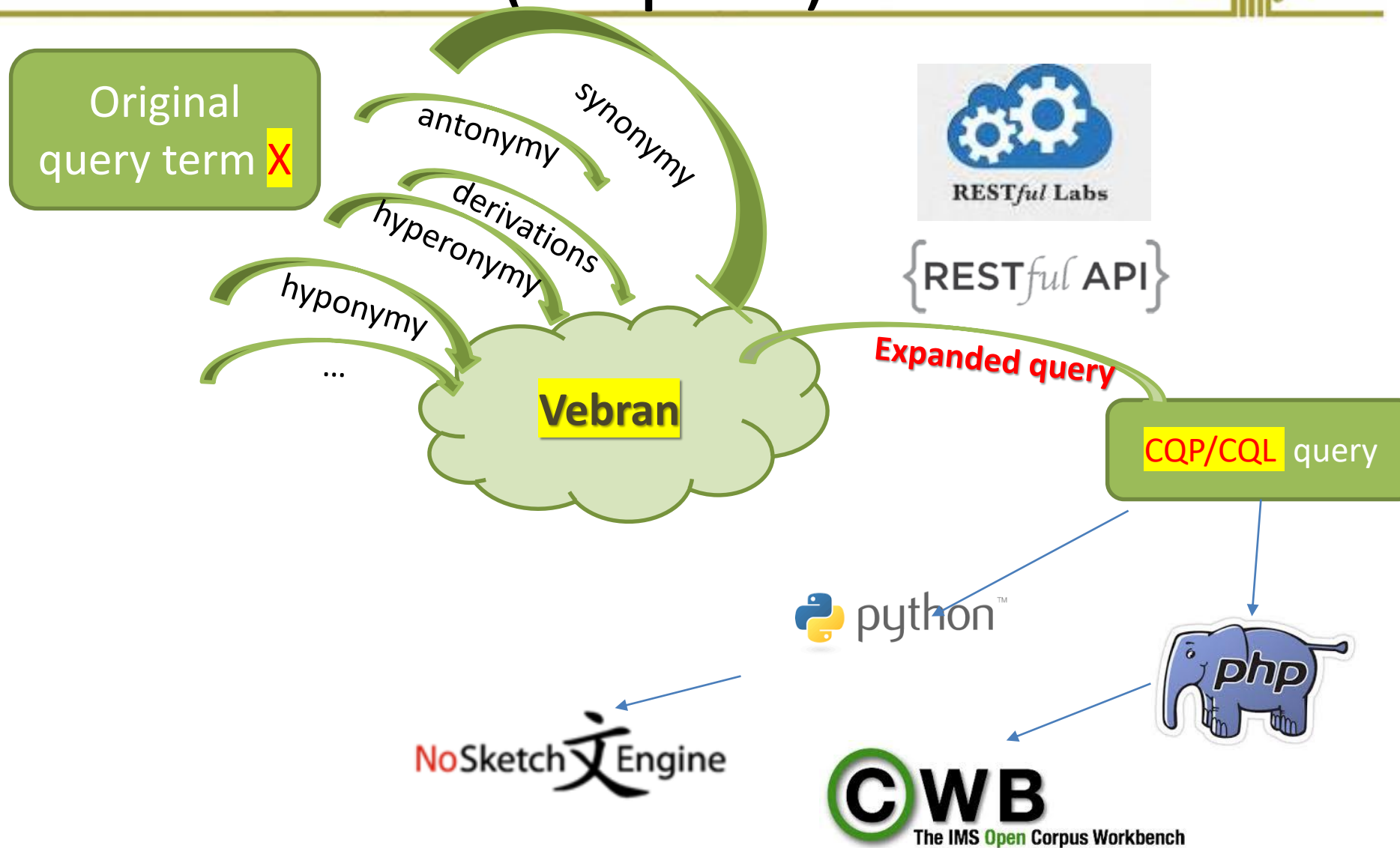
Implementation

- Vebran, an adapted set of web services (RESTfull, MVC .Net)
- Modification of the search interfaces' open source code (PHP, Python)

How do we expand (digital libraries)?

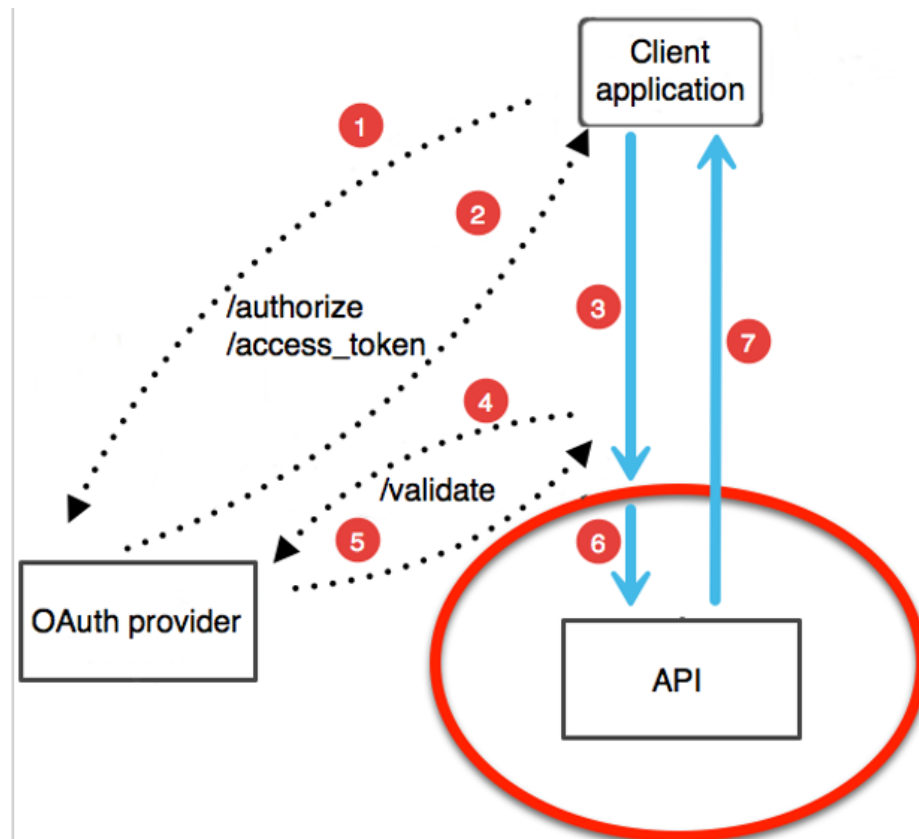


How do we expand (corpora)?





- Query expansion
 - Delaf/obliciZaCQP
 - Delafs
 - Sinonimi
 - Semrel
 - Query
- In progress other:
 - textCorrect
 - NER



OAuth 2.0 Access Token Enforcement
Using External Provider policy

Morphological expansion



- Idea: expand the original query, in which lemma X is given, with corresponding inflected forms in specified alphabet(s) and optionally with restrictions regarding grammatical categories.
- Inflected forms:
 - from Unitex delaf and delacf files
 - MWU inflected forms supported by rule based system

Semantic expansion



- Idea: expand the original query, in which term X is given, with other terms that are in a semantic relation with the term X (synonymy, antonymy, hyperonymy, etc.).
- Lexical relations
 - Wordnet: hypernym, hyponym, holo_member, holo_part, eng_derivative, near_antonym,...

Functions of .../vebran/api/



<http://hlt.rgf.bg.ac.rs/vebran/api/delaf/>

Input: "{lema: 'sreća', alphOut:'C', lngIn:'sr', lngOut:'sr',
POS:'N', GramCats:'p', fleksije:false, dlfByLemma:false}";

Returns: **cpeñ(a | ama | e);**

<http://hlt.rgf.bg.ac.rs/vebran/api/obliciZaCQP/>

Input: lema, ?POS

Returns: **srecx(a | ama | e | i | o | om | u);**

.../vebran/api/ functions



<http://hlt.rgf.bg.ac.rs/vebran/api/delafs/>

Input: 'sreća' (options lemma and optional POS)

Returns: **sreća;srećama;sreće;sreći;srećo;srećom;sreću;**
срећа;срећама;среће;срећи;срећо;срећом;срећу

<http://hlt.rgf.bg.ac.rs/vebran/api/sinonimi/post>

jsonLema = "{lema:'sreća', alphOut:'C', lngIn:'sr', lngOut:'sr',
POS:'N', GramCats:'s', fleksije:false, dlfByLemma:true}";//

Returns: **S:околности;S:срећа | срећама | среће | срећи | срећо | срећом | срећу;**
S:судбина | судбинама | судбине | судбини | судбино | судбином | судбину

<input checked="" type="checkbox"/> WordNet...	chance, circumstances, destiny, fate, felicity, fortune, happiness, hazard, lot, luck, portion	↑ ↓	okolnosti, sreća, sudbina
--	--	--------	---------------------------

.../vebran/api/ functions



<http://hlt.rgf.bg.ac.rs/vebran/api/semrel/post>

```
jsonLema = "{relType: 'derived', lema:'režimski', alphOut:'LC', ...}";
```

Returns: **S:vlast;S:režim;S:vlada;S:власт;S:режим;S:влада**

```
jsonLema = "{relType: 'near_antonym', lema:'mlada osoba', ...}";
```

Returns: **C:odrastao čovek;C:zreo čovek;C:одрастао човек;C:зрео човек**

Relation types: *hyponym, hypernym, holo_member, mero_member, eng_derivative, mero_portion, holo_part, category_domain, near_antonym, derived, be_in_state, holo_portion, also_see, similar_to, verb_group, region_domain*

SrpKor: CQP query without morphological expansion



- SrpKor has been automatically PoS-tagged and lemmatized using TreeTagger. The ambiguity of forms with short-length lemmas, e.g. “tat” (*thief*) and “tata” (*dad*), causes tagging errors (almost all forms of “tata” are tagged as forms of “tat”). As a result, query

[lemma=“tata”]

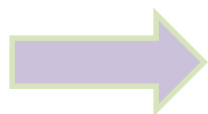
gives only 30 results and only word types “tatama” (poor recall). Actually, “tata” is far more frequent than “tat”.

SrpKor: CQP query with morphological expansion



- We introduce fake positional attribute **flemma**

[flemma="tata"] (User query)



(Vebran expands query using DELA)

[word="tat(a|ama|e|i|om|u)"]

- User query is replaced with regular CQP syntax before sending it to the CQP interpreter.
- Now we have 2171 results (100% recall), but still, due to homographs, it's possible that some retrieved forms don't correspond to the given lemma.
- Actually, [flemma="tat"] would produce similar results, most of them would not be relevant.

RudKor (NoSketch Engine)



Query **kap(a|ama|e|i|o|om|u)** 69 (19.48 per million) **i**

Page of 4 [Next](#) | [Last](#)

doc#6	obezbeđuju laminarno strujanje . Zaostale	kapi	nafte i čvrste suspen
doc#28	, ispitivanja kvaliteta vode , izrade zaštitne	kape	bunara , likvidacije r
doc#28	privremene zaštitne kape bunara . Zaštitna	kapa	bunara se izrađuje oc
doc#28	i nadfilterske (pune) cevi . Spajanje zaštitne	kape	bunara i nadfilterske
doc#28	u zasipu takođe se postavlja zaštitna	kapa	. Pijezometarska kap
doc#28	se postavlja zaštitna kapa . Pijezometarska	kapa	je izrađena od pocinl
doc#28	obezbeđenju , bunarskoj i pijezometarskoj	kapi	, kao i o likvidaciji ra
doc#31	nukleusi za nastanak većih agregata kao što su	kapi	vode (magla , kiša)

- Similar example: “kap” (*a drop*) vs. “kapa” (*a cap*). We looked for “kapa”, but the first and the last result correspond to lemma “kap”.

Expansion of CQP query using synonyms (SrpKor)



- Now we introduce fake positional attribute **synlemma**.
Vebran API uses only Wordnet and terminological databases to extract lemmas of synonyms, but there is an option to extract forms of synonyms as well using DELA.

[synlemma="srecxa"]



[lemma="okolnosti|srecxa|sudbina"]

- (Remark for the editors of Unitex dictionaries and Serbian Wordnet: “okolnosti” exists in WordNet as a lemma (synonym for “sreća”) and only as a plural form of lemma “okolnost” in DELA dictionaries, but not as a (pluralia tantum) lemma. So, there won’t be any forms of “okolnosti” in search results.)

34. poli071127.txt:

je obelodanjeno da su se trojica napadača sakrila u žbunje u blizini pružnog prelaza i tu pripremila zasedu policijskom vozilu . Strpljivo su čekali da bi pucali u policajce koji su bili u njemu . Na [sreću](#) , plan terorista nije realizovan . Pošto ni sat vremena posle pripremljene zasede , policijsko vozilo nije prošlo predviđenom maršrutom , napadači su otišli kućama . S obzirom na to da Tači , Rama i

38. 045 2458-04.txt:

arstva (u daljem tekstu : ovlašćeno lice) . Priređivači su dužni da omoguće ovlašćenom licu pregled prostorija i uvid u radnje koje su , neposredno ili posredno , povezane sa priređivanjem igara na [sreću](#) , poslovne knjige , izveštaje , evidencije , softvera i druga dokumenta ili podatake , na osnovu kojih se može utvrditi poslovanje priređivača . Ovlašćeno lice može prisustvovati otvaranju , obračuna

50. mm.xml:

ge od farbanog patosa pa sve do čađave tavanice , i peć . Ivan je doznao da su gost i njegova tajna žena već prvih dana svoje veze zaključili da ih je na uglu Tverske ulice i one uličice spojila sama [sudbina](#) i da su njih dvoje stvoreni jedno za drugo zauvek . Ivan je doznao iz pripovedanja gosta kako su zaljubljeni provodili dan . Ona bi došla i , kao prvo , nadevala kecelju u uskom predsoblju , gde se

51. Dosljaci-sve.xml:

, mnogo patila , ali sam i mnogo ljubila . Malo je žena koje je ljubav tako slatko ljuljkala kao mene . Ona je obuzimala najviše vrhove moje misli . I ja mogu reći s pravom : ' Hvala ti , Bože , moja [sudbina](#) je bila dobra . ' Ne pati misleći da mi je smrt bila teška . Smrt je prosta . Zašto je se bojati ? Ona je kao nesvest koja obuzima one koji su dostigli granicu napora . Ja sad vidim koliko se nepravo

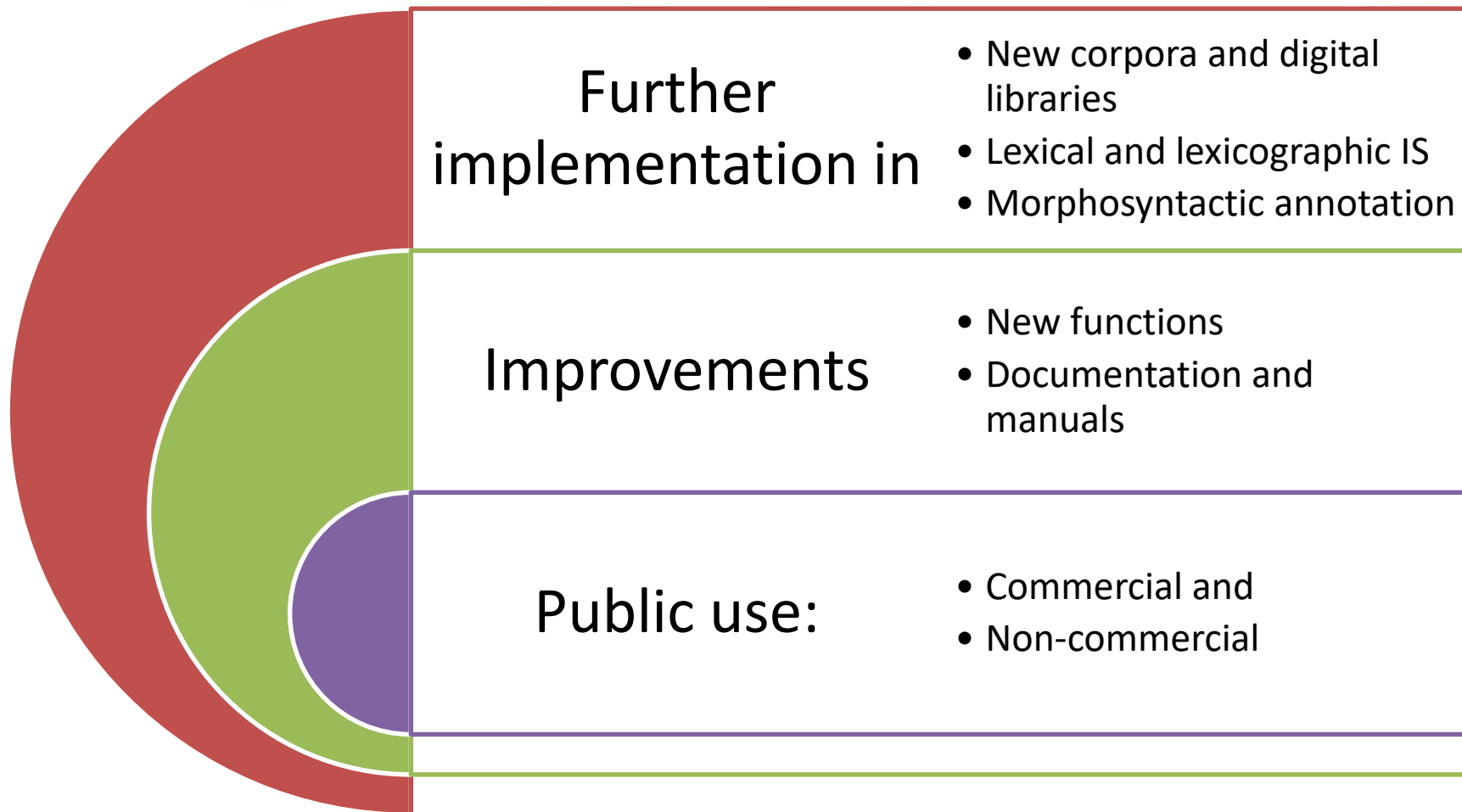
Bibliša - bilingual



SEARCH RESULTS

Number of concordances (en/de/fr): 77		Broj konkordansi (sr): 77
Vasiljević, 2013, vol. XIV:1, ID: 1.2013.1.7 metadata	We can say that Python encompasses the key advantages of all of these programming languages , regarding learning to program:	Možemo reći da Pajton obuhvata glavne dobre osobine svakog od ova tri programaska jezika , kada je u pitanju učenje programiranja:
Vasiljević, 2013, vol. XIV:1, ID: 1.2013.1.7 metadata	This is the same as when people were beginning to learn higher programming languages , before they learned assembler and implementation of language processors.	To je isto kao što su u nekom trenutku ljudi počeli da uče više programske jezike , a da nisu morali prethodno da nauče assembler i implementaciju jezičkih procesora.
Vasiljević, 2013, vol. XIV:1, ID: 1.2013.1.7 metadata	The Pascal programming language supports nested functions (a function defined within another function), but despite the fact that Pascal has influenced almost all subsequent programming languages in terms of the elements of structured and procedural programming, many programming languages do not support nested functions.	Programski jezik Paskal (engl. Pascal) podržava ugnježdene funkcija (definisanje funkcije unutar druge funkcije), ali i pored toga što je Paskal uticao na gotovo sve kasnije programske jezike kada su u pitanju elementi strukturiranog i proceduralnog programiranja, mnogi programski jezici nemaju podršku za ugnježdene funkcije.
Stanković et al., 2011, vol. XII:1, ID: 1.2011.1.4 metadata	The application was developed using MS Visual Studio and the programming languages C# and AspX, while the data were stored in the MS SQL Server 2008 data management system.	Za razvoj aplikacija korišćen je MS Visual Studio i programski jezici C# i AspX, a podaci su skladišteni u MS SQL Server 2008 sistemu za upravljanje podacima.
Nastić, 2008, vol. IX:1/2, ID: 1.2008.1/2.8 metadata	Most often those courses were introductory courses in Computer Science such as Basics of Computer Systems, Programming Systems, Programming Languages , Computing Machines and Programming, Application of Computers and Introduction to Cybernetics.	To su najčešće biliprvi kursevi iz oblasti računarstva, kao što su: Osnovi računarskih sistema, Programski sistemi, Programski jezici , Računske mašine i programiranje, Primena računara i Uvod u kibernetiku.

Also planned...





ranka@rgf.rs
misko@matf.bg.ac.rs