

# Vebran Web Services for Corpus Query Expansion

UDC 004.738.52:811.163.4'373

DOI 10.18485/infodhca.2019.19.2.5

**ABSTRACT:** This paper discusses the development of the Vebran web services and their application to corpus search improvements. The Vebran web services are used to consult external lexical resources for Serbian (mainly electronic morphological dictionaries and Serbian Wordnet) and expand user queries to retrieve more relevant results from Serbian corpora.

**KEYWORDS:** corpus search, web service, Serbian lexical resources, query expansion.

**PAPER SUBMITTED:** 18 October 2019

**PAPER ACCEPTED:** 13 December 2019

Ranka Stanković

ranka.stankovic@rgf.bg.ac.rs

University of Belgrade

Faculty of Mining and Geology

Miloš Utvić

milos.utvic@fil.bg.ac.rs

University of Belgrade

Faculty of Philology

Belgrade, Serbia

## 1 Introduction

It is generally known that users typically formulate very short instead of long and complex queries. The lemmatization of corpora enables easy retrieval of all inflected forms for a lemma, but one can not be sure in the accuracy of automatic lemmatization.

In order to achieve retrieval effectiveness as regards these short queries, we need special techniques, because a straightforward keyword match may not always be adequate. The main objectives are 1) to upgrade the existing web interfaces for searching through language resources and 2) to enable querying language resources supported with available lexical resources.

The language resources to be searched are various digital libraries and corpora, but in this paper, we will focus on corpus case study. The query expansion will rely on different lexical resources to support the search: 1) morphological electronic dictionaries; 2) WordNets and 3) terminological databases. The new web services named Vebran are developed as an upgrade of the previous service wsQueryExpand (Stanković, 2009; Stanković et al., 2012). The Vebran is implemented as Restful API with several new functions, but the biggest improvement is a set of functions specially prepared to support query expansion based on lexical resources.

Sections 2 and 3 describe language resources for Serbian, corpora that we can search and lexical resources that Natural Language Processing (NLP) applications can consult. Vebran web services and their usage of lexical resources are discussed in Section 4. Morphological and semantic corpus query expansion, based on Vebran web services, is described in detail in subsections 4.2 and 4.3 respectively. Finally, in Section 5 we conclude with some remarks concerning the future work.

## 2 Corpora

Numerous corpora are developed within research activities of Human Language Technology (HLT) Group at the University of Belgrade and the Language Resources and Technologies Society (JeRTeh):

- monolingual general corpora: Corpus of Contemporary Serbian (versions SrpKor2003 and SrpKor2013)<sup>1</sup> and its subset SrpLemKor<sup>2</sup>;
- SrpEngKor<sup>3</sup>, aligned English-Serbian corpus including subcorpus SELFEB (Serbian-English Law Finance Education and Health) with documents on finance, health, law and education;
- SrpFranKor<sup>4</sup>, aligned French-Serbian corpus;
- SrpNemKor<sup>5</sup>, aligned German-Serbian corpus;
- RudKor<sup>6</sup>, a specialized monolingual corpus of texts from the mining domain, etc.

The query expansion will be demonstrated using examples in two Serbian corpora: SrpKor2013 (cf. 2.1) and RudKor (cf. 2.2). Both corpora are morphologically tagged in the sense that each token is associated with the information about the corresponding part of speech and lemma as described in (Утвић, 2011, 2014). The TreeTagger software tool (Schmid, 1997, 1999) was used for automatic morphological annotation of both corpora. The Tree-Tagger language parameter file for Serbian was created as a derivative of a system of Serbian Morphological electronic Dictionaries (SMD, cf. 3.1), authored by Cvetana Krstev and Duško Vitas (Krstev, 2008).

---

<sup>1</sup> <http://www.korpus.matf.bg.ac.rs>

<sup>2</sup> <http://www.korpus.matf.bg.ac.rs/SrpLemKor/>

<sup>3</sup> <http://www.korpus.matf.bg.ac.rs/SrpEngKor>

<sup>4</sup> <http://www.korpus.matf.bg.ac.rs/SrpFranKor>

<sup>5</sup> <http://jerteh.rs/biblisha/>

<sup>6</sup> <http://147.91.181.179/cqp/cqpweb>

## 2.1 SrpKor2013

The Corpus of Contemporary Serbian (SrpKor) was established in 2002, at the initiative of Professor Ljubomir Popović, with the aim of enabling researchers to consult the chosen collection of Serbian texts via the Internet (Vitas and Krstev, 2012). The first version of SrpKor, SrpKor2003, has not been morphologically annotated.

SrpKor2013 is the current version of SrpKor (Utvić, 2014), used as a reference and general purpose corpus containing over 122 million corpus words. It includes literary texts of Serbian writers in the XX and XXI centuries, as well as scientific and popular science texts from different domains (natural and social sciences), administrative and general texts. The general texts represent articles from the daily newspapers “Politika”, “Večernje Novosti”, “Danas”, texts from magazines “Danica”, “Ebit”, “Ekonomist”, “Glasnik”, “NIN”, “Ilustrovanja politika”, “Kalibar”, “Moje srce”, “Mostovi”, “Pravoslavlje”, “Svet”, “Teološki pogledi”, “Trn”, “Viva”, “Republika”, texts from the internet portal “Peščanik”. Some of the texts are translations, most of which are literary texts, while a smaller part are translations of general texts. Apart from being morphologically annotated, the corpus texts are provided with corresponding bibliographic description, information concerning the functional style to which the text belongs, as well as an indicator whether corpus text is written in Serbian or represents a translation from another language.

The SrpKor2013 is not structurally annotated, although some or all levels of the text structure (section, title, paragraph, sentence) are annotated in some particular corpus texts, especially those which are part of aligned corpora.

The SrpKor2013 corpus is used by more than 700 users, mostly Slavists.

## 2.2 RudKor

Systematic collection and preparation of texts from the mining domain started with English-Serbian alignment of articles in a bilingual journal “Podzemni radovi”, followed by mining projects, law regulations, PhD theses and textbooks from the mining domain. Texts are gathered and organized in the ROMeka@RGF digital library (Tomašević et al., 2018). The RudKor corpus originated from ROMeka@RGF digital library to enable various linguistic and terminological research, including extraction of terms and other tasks in the field of knowledge engineering (Утвић et al., 2018). The RudKor contains 344 different texts with a total size of 5.4 million words.

Mining terminology is introduced in the system of Serbian Morphological Dictionaries (cf 3.1). In order to allow the extraction of specific concepts and relations between concepts by creating lexical masks, new semantic markers relevant to the field of mining have been integrated (Обрадовић et al., 2017).

### 2.3 Corpora tools

Three different systems for diverse types of usage scenarios are used in this research:

- Unitex (Paumier, 2016; Krstev, 2008);
- Open Corpus Workbench (OCWB) (Evert and The OCWB Development Team, 2019) and web-based graphical user interface CQPweb (Hardie, 2012);
- NoSketch Engine (Rychlý, 2007).

Unitex<sup>7</sup> is open source software for an analysis of textual data, corpus processor with user-friendly interface, language resources distributed out-of-the-box and set of functions that can be used from other software systems. The Unitex NLP engine is based on automata-oriented technology, allowing users to 1) Compile rules and dictionaries as finite-state machines; 2) Use variables instanced with a part of the text or with any characters; 3) Use regular expressions and graphs of automata and transducers for searching and extraction; and 4) Build cascades of rules.

Corpora SrpKor2013 (cf. 2.1) and RudKor (cf. 2.2) can be searched by OCWB, while a search of RudKor is also available through NoSketch Engine.

This paper presents the process of upgrading the existing corpus search web interfaces of OCWB and NoSketch Engine in order to enable corpus query expansion.

## 3 Lexical resources

In order to improve the current corpus search capabilities based on linguistic annotation, it is necessary to consult external lexical resources. The following lexical resources have been developed for Serbian by the HLT Group at the University of Belgrade and JeRTeh Society:

- System of Serbian morphological electronic dictionaries (Unitex DELA format);

---

<sup>7</sup> <https://unitexgramlab.org/>

- Semantic network WordNet for Serbian;
- Terminological databases Termi, RudOnto, GeolISS.

### 3.1 Serbian morphological resources

The system of morphological electronic dictionaries of the Serbian language (SMD) (Krstev, 2008) is the core for the morphological expansion. SMD follows the methodology and format (known as DELAS/DELAF) that was developed in LADL (Laboratoire d'Automatique Documentaire et Linguistique) under the guidance of Maurice Gross. The format of a DELAS-type dictionary basically consists of simple word lemmas, each accompanied by inflectional class code. Every inflectional class code is associated with a corresponding finite-state transducer responsible for the generation of all inflectional forms of DELAS lemma. Thus, the DELAS-type dictionary with finite-state transducers for inflection enables the production of a DELAF-type dictionary which consists of all inflectional forms of DELAS-lemmas with their corresponding grammatical information. The Serbian morphological dictionary of simple words contains 190,000 lemmas which yield the production of approximately 2.4 million different inflected forms for lemmas and about 7.6 million forms with associated grammatical categories. At present, the dictionary of compounds has about 18,000 lemmas covering different parts of speech.

Lexical data have been migrated from textual e-dictionaries to a lexical database. After years of development, SMD, developed as a system of textual files, have become a large and complex lexical resource. An on-line application for dictionary development and management, based on a central lexical data repository (lexical database) is developed offering various possibilities for improvement of SMD, e.g. control of data consistency and introduction of explicit relations between lexical entries, automatic generation of dictionary candidates. The new version of service Vebran (cf. 4) is using this database for morphological expansion (Stanković et al., 2018).

The automatic procedure was used to transfer data from the existing dictionaries into the lexical database and to store all information about lemma and form entries as structured data. DELAS-lemma entries are generally mapped to entries in tables `LexicalEntry` and `LexicalSense` (Figure 1). A lemma, its corresponding PoS and inflectional class code (defining all inflected forms) are stored in the `LexicalEntry` table, while associated syntactic, semantic, domain and other types of markers are separated. Identical lexical entries from DELAS sharing the same inflectional class (e.g. `vez,N297`)

are merged into one `LexicalEntry`, while associated markers that differentiate senses are recorded in the `LexicalSense` and `SenseProperties` tables.

All inflected forms of lemma `vez` (`vez`, `veza`, `vezu`, `veze`, `vezom`, `vezovi`, `vezova`, `vezovima`, `vezove`) are stored in the table `Forms`, together with sets of grammatical categories assigned. Since one lexical form can represent one or more grammatical realization of a lexical entry, it is described with one or more sets of grammatical categories stored in `FormGramCats` table. For instance, the form `vezom` has one set of grammatical categories assigned to it `:ms6q` (the instrumental case, singular), while three sets of grammatical codes (`:ms2q`, `:mw2q`, `:mw4q`) are assigned to the form `veza` (the genitive case, singular and paukal, as well as accusative paukal). In addition, sets of grammatical categories are represented as individual categories in the table `FormGramCatProperties`, as presented on the left side of Figure 1. More details about multi-word units mapping, markers and relations between lexical entries can be found in (Stanković et al., 2018).

### 3.2 Semantic networks

The Serbian wordnet (SWN) has been developed in the scope of the Balkanet project following the model adopted for the EuroWordnet project (Krstev et al., 2004). More than 22,000 synsets built by app. 28,000 literals are Princeton WordNet (PWN), except for 532 Balkan specific concepts that are connected with other Balkan languages, and 155 Serbian specific concepts that remain unconnected with other languages. Since the core of wordnets developed for Balkan languages was produced by translation of the basic synsets in the PWN 3.0, the hypernym/hyponym relations in SWN mirror its hierarchical structure. Other relations are implemented more freely, depending on specific lexicalizations in Serbian. These relations include antonymy, meronymy, as well as some cross-part of speech relations (XPoS), such as causes and `be_in_state`.

A synset `ENG30-14473222-n` (Figure 2), visualized by Hydra<sup>8</sup> (Rizov and Dimitrova, 2016), defined as *Nečije opšte okolnosti ili uslovi u životu (uključujući sve što vam se dešava)* (“your overall circumstances or condition in life (including everything that happens to you)”) has three literals: *okolnosti*, *sudbina* and *sreća* (“circumstances”, “destiny”, “luck”). An example of hypernym to `ENG30-14473222-n` is *uslov* (“condition”) and the corresponding hyponym is *srećne okolnosti* (“lucky circumstances”). All mentioned relations can be used for query expansion.

<sup>8</sup> <https://dcl.bas.bg/bulnet/>

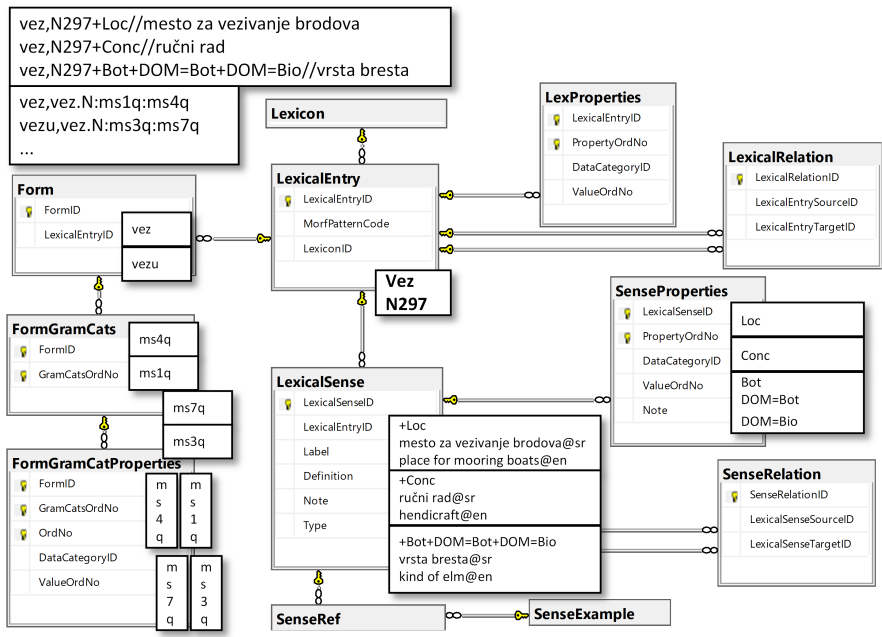


Figure 1. DELA to lexical database mappings

Figure 2. Wordnet synsets with literal “sreća”

### 3.3 Terminological resources

Termi<sup>9</sup> application supports the development of terminological dictionaries in various fields (mathematics, agronomy, mining), as well as the processing and presentation of terms in Serbian and English. Termi has been recently supplemented by the general Serbian-German bilingual dictionary extracted from 14 contemporary novels (Andonovski et al., 2019). All verified terms are available in public view, while additional terms are internally available to authorized users, according to their role and domain. A hierarchical display of the vocabulary terms is available for each domain. Besides its name, each term has its synonyms, abbreviations, description and bibliography. In case that the description of a term contains a L<sup>A</sup>T<sub>E</sub>X fragment, the fragment will be interpreted, which helps in the presentation of mathematical formulae.

Rudonto is a terminological resource developed to support knowledge management in mining engineering, focusing on the application in mining equipment and mine safety domains. Through export to several specific formats, RudOnto ontologies offer the possibility of generating stand-alone terminological resources or ontologies from specific sub-fields (sub-domains) (Kolonja et al., 2016).

GeolISSTerm represents the core of GeolISS (GEOLogical Information System of Serbia), and it is implemented as an aggregation of geological vocabularies, collections of terms and text definitions of entities thought to exist in a domain or collections of possible values for properties. The terms in the vocabularies are used to classify observations/interpretations, or to specify attribute values. GeolISSTerm is organized as a taxonomy with definitions for each entry, accompanied by synonyms, bibliographical references, equivalent terms and definition in another language (presently only English equivalents of definitions are in the database).

Externally developed Dictionary of library and information sciences<sup>10</sup> encompasses the terminology of theory and practice of librarianship and information sciences and a wide range of close or related fields, in Serbian, English and German languages. The languages in this dictionary have equal status. Online version currently includes 40,000 entries, but for the implementation of our web service, the older version is used with 23,400 entries (11,300 in English and 12,100 in Serbian, 910 definition or annotation terms which belong to library standards, and 2,200 acronyms of international and national entities). The intention of this dictionary is to be the useful elec-

---

<sup>9</sup> <http://termi.rgf.bg.ac.rs/>

<sup>10</sup> <http://rbi.nb.rs>



tronic resource of information for Library and Information Science professionals, for scientists and students, as well as for library users with different interests.

## 4 Vebran Web Services

Vebran Web Services enable users to search corpora using query syntax which is not supported by back-end query processors of OCWB (CQP) and NoSketch Engine (Manatee) in the following way:

- user can request that particular term *X* in a given query should be replaced with lemmas or word forms of terms which are semantically related to *X* in some manner (synonyms of *X*, antonyms of *X*, meronyms of *X*, hyponyms of *X*, hypernyms of *X*);
- user can request that particular lemma *X* in a given query should be replaced with its inflectional paradigm, i.e. with all word forms of lemma *X*.

### 4.1 Services architecture

The architecture of query expansion via Vebran Web Services (Figure 3) resembles a typical client-server architecture. Corpus web search interfaces, OCWB/CQPweb and NoSketch Engine / Bonito, perform the role of clients for Vebran Web Services, requesting a set of lemmas or word forms related to a given term *X*.

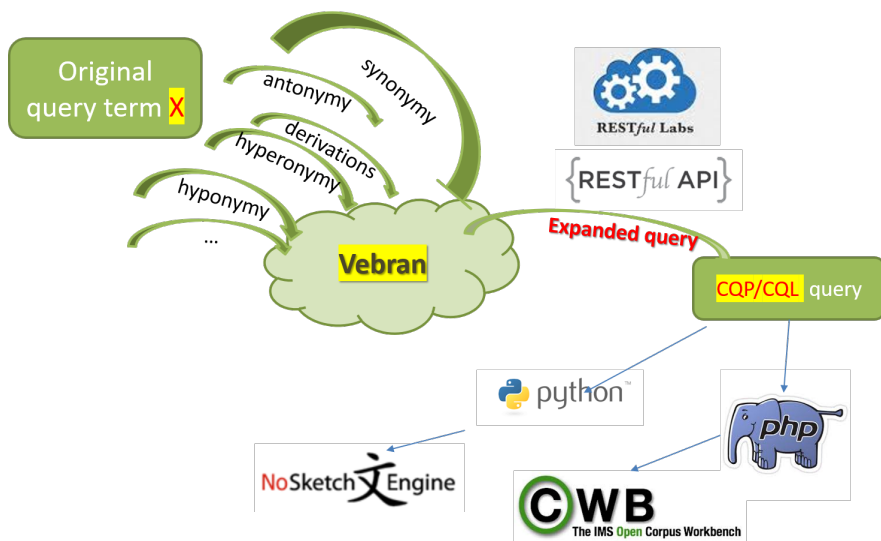
Firstly, clients need to get authorization in order to send their requests to Vebran Web Services. In this case, clients use an access token to identify themselves (Figure 4). The only parties that should ever see the access token are the client itself, the authorization server, and the resource server. The client should ensure that the storage of the access token is not accessible to other clients on the same device. The access token can only be used over a https connection, since passing it over a non-encrypted channel would make it trivial for third parties to intercept. The token endpoint is where apps make a request to get an access token for a user.

After successful authorization, clients are allowed to send a request specifying the term *X* and the relation (semantic or morphological) which should exist between *X* and the requested lemmas or word forms. Based on the client's request, Vebran services consult external lexical resources (see Section 3) and generate a response to the client. Communication with Vebran Web Services is based on RESTful technology, implemented in the Microsoft

MVC.Net web application framework which uses the model–view–controller pattern.

Clients are open source software (OCWB/CQPweb and NoSketch Engine / Bonito have been implemented in PHP and Python respectively) and their source code has been adapted to:

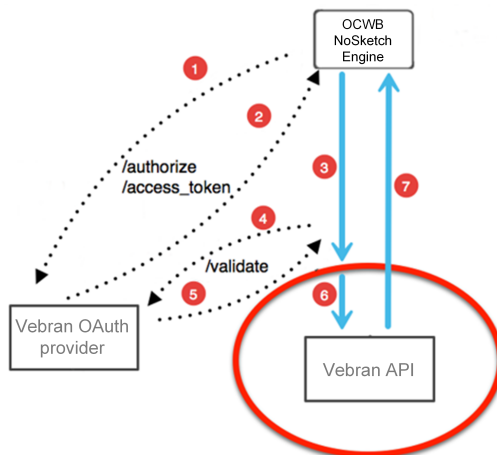
- send requests to Vebran Web Services,
- receive a response from Vebran Web Services,
- expand a given user query with a response from Vebran Web Services producing a new query with syntax acceptable by back-end query processors (OCWB/CQP and NoSketch Engine / Manatee).



**Figure 3.** Query expansion architecture

## 4.2 Morphological expansion

Morphological query expansion is an alternative to queries based on a part-of-speech annotation of the corpus. That alternative is necessary to



**Figure 4.** Oauth 2.0 Access Token Enforcement

overcome recall problems caused by tagging errors and limitations imposed by the format of a TreeTagger full-form lexicon. Each entry of the TreeTagger full-form lexicon contains one-word form and a sequence of tag-lemma pairs that could correspond to that word form (Schmid, 1997). TreeTagger full-form lexicon does not allow the possibility of a lexicon entry with two or more tag-lemma pairs corresponding to the same word form and having the same PoS tag, while having different lemmas. However, it is a common case that different short-length lemmas in Serbian, e.g. nouns *tat* (“thief”) and *tata* (“dad”) have homograph word forms (*tati*, *tatom*, *tate*, *tatu*, *tata*) causing that lexicon entries with these forms cannot contain both tag-lemma pairs (N, *tat*) and (N, *tata*) where N is PoS tag denoting noun. Thus, creator of full-form lexicon has to choose which tag-lemma pair will keep and the choice is commonly based on the automatic process which randomly favors one lemma over another. As a result, the query [lemma="tata"] in SrpKor2013, automatically PoS-tagged by TreeTagger, gives only 30 results and only word types *tatama* (poor recall). Actually, forms of lemma *tata* are far more frequent in Serbian (SrpKor2013, too) than forms of lemma *tat*.

OCWB (Evert and The OCWB Development Team, 2019) and NoSketch Engine (Rychlý, 2007) treat corpus as a table where :

- each row represents either a particular corpus position (token) or an XML structure tag;
- only the first column (called **word**) is mandatory and represents token values or XML tags;
- if the table includes more columns then each column represents a specific type of information (linguistic and non-linguistic) associated with a corresponding token.

The names of corpus columns, also called *positional attributes* by OCWB and NoSketch Engine, are used in queries in the form

```
[positionalAttribute="regularExpression"].
```

Besides the column **word**, SrpKor2013 also uses positional attributes **pos** (part of speech) and **lemma**, while RudKor uses positional attributes **tag** (part of speech) and **lemma**.

The general idea behind the morphological expansion is to replace lemma X in a given user query with the corresponding inflected forms of X in the specified alphabet(s) and, optionally, with restrictions regarding grammatical categories. Inflected forms are stored in LeXimirka database and originate from Unitex DELAF and DELACF dictionaries, described in Section 3.1. The inflection of multiword units is additionally supported by the rule based system. The system supports different alphabets and character encodings (the aurora alphabet and ISO-8859-1 character encoding for SrpKor2013 corpus, Serbian Latin alphabet and UTF-8 character encoding for RudKor corpus), in order to enable query expansion for different corpora.

There are several functions of Vebran Web Services which handle the morphological expansion:

- **delaf**<sup>11</sup>,
- **obliciZaCQP**<sup>12</sup>,
- **delafs**<sup>13</sup>.

Function **delaf** expects input parameters (Table 1) via a POST request as a JSON (JavaScript Object Notation) structure like Figure 5 representing all plural word forms of lemma/noun **sreća** (“happiness”) in Serbian using Serbian Latin alphabet. The appropriate output result is in the form of a regular expression: **sreć(a|ama|e)**, that can effectively retrieve all inflected plural forms as requested.

---

<sup>11</sup> <http://hlt.rgf.bg.ac.rs/vebran/api/delaf/>

<sup>12</sup> <http://hlt.rgf.bg.ac.rs/vebran/api/obliciZaCQP/>

<sup>13</sup> <http://hlt.rgf.bg.ac.rs/vebran/api/delafs/>

```

{
  lema: 'sreća',
  alphOut: 'L',
  lngIn: 'sr',
  lngOut: 'sr',
  POS: 'N',
  GramCats: 'p',
  fleksije: false,
  dlfByLemma: false
}

```

**Figure 5.** JSON structure of a request for all plural word forms of lemma/noun *sreća* (“happiness”) in Serbian using Serbian Latin alphabet

Parameter	Value examples	Description
lema	sreća	Requested lemma X
alphOut	C	Alphabet of the output result C-Cyrillic, L-Latin, A-Aurora, combinations like CL, CA, LA, CLA are allowed
lngIn	sr	language of a given lemma X, by default sr (Serbian)
lngOut	sr	language of the function result, by default sr (Serbian)
POS	N	part of speech for a given lemma X
GramCat	p	morphological constraints as SMD data category values: s-singular, p-plural, 1..7 — case restrictions, etc.
fleksije	false	indicator whether the result contains only lemmas X or their inflectional forms as well (true, false)
dlfByLemma	false	indicator of output format (forms grouped by lemma or not)

**Table 1.** Parameters for morphological expansion requests (function `delaf`)

Function `obliciZaCQP` requires lemma as an input parameter, while part of speech is optional. The function is adapted to `SrpKor` which uses `aurora` alphabet. The function result is a regular expression (CQP and Manatee syntax) using `aurora` alphabet. The example (Figure 5) with `alphOut:'A'` would return `srecx(a|ama|e|i|o|om|u)`.

Function `delafs` uses the same input parameters, but generates output in the form of a list: `sreća; srećama; sreće; sreći; srećo; srećom; sreću; sreña; sreñama; sreñe; sreñi; sreño; sreñom; sreñy`. This format is used for query expansion in digital libraries `Romeka` and `Bibliša`, since their query processors require such forms.

Web search interfaces for `OCWB` and `NoSketch Engine`, `CQPweb` and `Bonito` respectively, have been adapted to accept and preprocess user query with an expanded syntax which is not supported by `OCWB` and `NoSketch Engine` corpus search engines. For morphological expansion fake positional attribute `flemma` is introduced allowing a user to request the inflectional paradigm of a lemma, e.g. `tata`, with a query `[flemma="tata"]`. Preprocessing includes:


- extraction of `flemma` value (e.g. `tata`);
- sending a request to `Vebran Web Services` (similar to Figure 5);
- using `Vebran Web Services` response to generate the final query (e.g. `[word="tat(a|ama|e|i|om|u)"]`) adjusted with allowed syntax of the query processor and
- sending the final query to the query interpreter.

In case of `[flemma="tata"]` vs. before used `[lemma="tata"]`, we get 2,171 results in `SrpKor2013` (100% recall) instead of earlier 30 results. However, due to homographs, it is possible that some retrieved forms do not correspond to the given lemma. Actually, `[flemma="tat"]` would produce similar results, but most of them would not be relevant.

A similar example can be found in `RudKor` through `NoSketch Engine` for lemma: `kap` (“a drop”) vs. lemma `капа` (“a cap”). The example page of search results for `капа` (Figure 6), shows that the first and the last concordance line correspond to lemma `kap`.

Although query expansion produces 100% recall, precision is reduced due to homographs.

The problems with query expansion in case of multi-word units (MWUs) described in (Утвић et al., 2019) have recently been resolved. Different solutions need to be applied for a case where all components of an MWU are the same as their lemmas, e.g. `leksički resurs` (“lexical resource”) and for

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

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

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

**Figure 6.** Expanded query in NoSketch Engine

a case where some components are not lemmas but instead some other inflected forms, like **leksička baza** (“lexical database”). For example, if the web service response contains the MWU **leksički resurs**, the generated query (`[lemma="leksički"] [lemma="resurs"]`) would appear to succeed because both the **leksički** and the **resurs** are lemmas of the corresponding lexemes. However, if the MWU **leksička baza** can be found in the web service response, the generated query `[lemma="leksička"] [lemma="baza"]` would find nothing since **leksička**, the inflected form of the lexeme **leksički**, is not a lemma of that lexeme.

Another problem is caused by the fact that MWUs may contain some components that inflect as a part of MWU and some that do not, e.g. **jato ptica** (“flock of birds”), where the first component inflects and the second does not (as a part of MWU), so generated query should be (`[lemma="jato"] [word="ptica"]`).

The solution for both problems includes web service function **MWUzaCQP**<sup>14</sup> which uses SMD to get information about MWUs, as well as an implementation of additional heuristics to process the out-of dictionary MWUs (Table 2):

1) If an MWU given by a user is found in SMD, its inflectional code is analysed and appropriate transformation applied. For instance, the inflectional code AXN associated with the MWU **leksički resurs** means that MWU contains three components, an adjective (A) that inflects, a separator (X) that does not inflect and a noun (N) that inflects. Another example is inflectional code N2X associated with MWU **jato ptica** where only the

<sup>14</sup> <http://hlt.rgf.bg.ac.rs/vebran/api/MWUzaCQP/>

first component (a noun) inflects. Web service answers client's request based on an inflectional code of MWU using markers **C**: (compound) **\_L** (lemma) and **\_W** (word), where the last two indicate whether marked string should be associated with the positional attribute **lemma** (it inflects as a part of MWU) or with the positional attribute **word** (it does not inflect as a part of MWU).

2) If SMD does not contain an MWU as a compound, each component of MWU is analysed separately. A component of MWU which has been found to be a lemma is associated with marker **\_L** (e.g. **ptica pevačica** in Table 2). If a component has been found to be an inflected form different from its lemma, component is then replaced with the corresponding lemma and marked with **\_L** (e.g. **leksička relacija** in Table 2). If a component cannot be found in SMD, it is treated as a word form that does not inflect (as a part of MWU) and it is marked with **\_W**.

Requested MWU	Inflectional code	Service response	Generated query
leksički resurs	AXN	C:leksički_L resurs_L	[lemma="leksički"] [lemma="resurs"]
leksička baza	AXN	C:leksički_L baza_L	[lemma="leksički"] [lemma="baza"]
jato ptica	N2X	C:jato_L ptica_W	[lemma="jato"] [word="ptica"]
ptica pevačica	?	C:ptica_L pevačica_L	[lemma="ptica"] [lemma="pevačica"]
leksička relacija	?	C:leksički_L relacija_L	[lemma="leksički"] [lemma="relacija"]
slobodan kao ptica	A3XN2	C:slobodan_L kao_W ptica_L	[lemma="slobodan"] [word="kao"] [lemma="ptica"]
album za slike	N4X	C:album_L za_W slike_W	[lemma="album"] [word="za"] [word="slike"]

**Table 2.** Examples of queries with MWUs

The presented heuristics to process queries with MWUs do not use recursive query expansion, that is, the inflectional paradigm of components is not produced in the form of a regular expression containing a



union of all word forms. Instead, the generated query uses positional attribute `lemma` and therefore a linguistic annotation provided by TreeTagger (last column of Table 2). An alternative would be to create a temporary query with expanded syntax, that is, with fake positional attribute `flemma` and then use a corresponding web service again as many times as there are components that inflect as a part of MWU. In that case, an example of the final query which searches for an inflectional paradigm of MWU `ptica pevačica` would be: `[word="ptic(a|ama|e|i|o|om|u)"] [word="pevačic(a|ama|e|i|om|u)"]`.

In the example `slobodan kao ptica` (“free as a bird”) the first and the last component inflect as parts of MWU, so the final query would be `[word="slobod(an|na|ne|ni|nih|nim|no|nog|noj|nom|nome|nu)"]`<sup>15</sup> `[word="kao"] [word="ptic(a|ama|e|i|o|om|u)"]`. The morphological category degree has been restricted to the value “positive” eliminating comparative and superlative forms from inflected forms of the adjective.

In example `album za slike` (“photo album”) only the first component has inflected forms: `[word="album(la|e|i|i|ima|om|u)"] [word="za"] [word="slike"]`.

An extra fake positional attribute `mwulemma` allows a user to request an inflectional paradigm of an MWU lemma, e.g. `leksički resurs`, with a query `[mwulemma="leksički resurs"]`.

### 4.3 Semantic expansion

The general idea to expand an original query with word forms related to term X is based on the use of semantic and terminological resources to find other terms such that there exists a given semantic relation (synonymy, antonymy, hyperonymy, meronymy) between those terms and the term X.

Web service function `sinonimi/post`<sup>16</sup> receives a JSON structure as an input and returns the synonyms of a given lemma. For example, function `sinonimi/post` for the lemma `sreća` returns `S: околности; S: срећа | срећама | среће | срећи | срећо | срећом | срећу; S: судбина | судбинама | судбине | судбини | судбино | судбином | судбину` as a set of corresponding synonyms and their inflected forms using Serbian Cyrillic alphabet.

More details for Semantic expansion can be found in (Утвић et al., 2019)

<sup>15</sup> Actually, this regular expression also includes `noga i nima`, but they are omitted to avoid text longer than length of line.

<sup>16</sup> <http://hlt.rgf.bg.ac.rs/vebran/api/sinonimi/post>

## 5 Conclusion

The paper describes Vebran web services that enable corpus query expansion and support corpus search that combines linguistic annotation of the corpus with external lexical resources. The presented approach allows search results to include the inflectional paradigm of the lexemes in a given user query, as well as the word forms semantically related to them (synonyms, antonyms, hyperonyms, etc.) where semantic relations are available through the semantic network (wordnet). The emphasis in this paper is on morphological query expansion supported by Vebran web services. Services consult external lexical resources and produce regular expressions that improve recall of the retrieved inflected forms for a given lemma. The described hybrid approach was successfully tested by modifying the web interface of corpus search tools OCWB and NoSketch Engine. Vebran web services are currently available only to the authorized users and applications. Further improvement of web services will include better support for multi-word units and a more flexible combination of different query parameters.

## Acknowledgment

This research was partly supported by the Ministry of Education, Science and Technological Development through projects ON-178006 and III 47003.

## References

- Andonovski, Jelena, Branislava Šandrih and Olivera Kitanović. “Bilingual Lexical Extraction based on Word Alignment for Improving Corpus Search”. *The Electronic Library* Vol. 37, no. 2 (2019): 722-739
- Evert, Stefan and The OCWB Development Team. *CQP Query Language Tutorial*, 2019, the IMS Open Corpus Workbench (CWB 3.4.16), May 2019. Accessed August 1, 2019. [http://cwb.sourceforge.net/files/CQP\\_Tutorial.pdf](http://cwb.sourceforge.net/files/CQP_Tutorial.pdf)
- Hardie, Andrew. “CQPweb — Combining Power, Flexibility and Usability in a Corpus Analysis Tool”. *International Journal of Corpus Linguistics* Vol. 17, no. 3 (2012): 380–409
- Kolonja, Ljiljana, Ranka Stanković, Ivan Obradović, Olivera Kitanović and Aleksandar Cvjetić. “Development of terminological resources for expert knowledge: a case study in mining”. *Knowledge Management Research & Practice* Vol. 14, no. 4 (2016): 445–456. <https://doi.org/10.1057/kmrp.2015.10>

- Krstev, Cvetana. *Processing of Serbian – Automata, Text and Electronic Dictionaries*. Belgrade: Faculty of Philology, 2008
- Krstev, Cvetana, Gordana Pavlović-Lažetić and Ivan Obradović. “Using Textual and Lexical Resources in Developing Serbian Wordnet”. *Romanian Journal of Information Science and Technology* Vol. 7, no. 1–2 (2004): 147–161
- Paumier, Sébastien. *Unitex 3.1 User Manual*, 2016, accessed August 1, 2019. <http://releases.unitexgramlab.org/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf>
- Rizov, Borislav and Tsvetana Dimitrova. “Hydra for Web: A Browser for Easy Access to Wordnets”. In *Proceedings of the Eighth Global Wordnet Conference, Research Institute for Artificial Intelligence, Romanian Academy*, 339–343, 2016
- Rychlý, Pavel. “Manatee/Bonito – A Modular Corpus Manager”. In *First Workshop on Recent Advances in Slavonic Natural Language Processing*, Sojka, P. and A. Horák, 65–70. Brno: Masaryk University, 2007
- Schmid, Helmut. “Probabilistic Part-of-Speech Tagging Using Decision Trees”. In *New Methods In Language Processing*, Jones, D. B. and H. Somers, Chapter 12, 154–164. Routledge, 1997
- Schmid, Helmut. “Improvements in Part-of-Speech Tagging with an Application to German”. In *Natural Language Processing Using Very Large Corpora*, Armstrong, S. et al. *Text, Speech and Language Technology*, Vol. 11, Chapter 12, 154–164. Dordrecht: Springer, 1999,
- Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić. “A Tool for Enhanced Search of Multilingual Digital Libraries of E-Journals”. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, (Istanbul, Turkey, 2012, 1710–1717
- Stanković, Ranka. “Modeli ekspanzije upita nad tekstuelnim resursima”. Phdthesis, Univerzitet u Beogradu, Matematički fakultet, Beograd, 2009
- Stanković, Ranka, Cvetana Krstev, Biljana Lazić and Mihailo Škorić. “Electronic Dictionaries - from File System to lemon Based Lexical Database”. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, McCrae, J. P., C. Chiarcos, T. Declerck, J. Gracia and B. Klimek, 48–56. Paris, France: European Language Resources Association (ELRA), 2018
- Tomašević, Aleksandra, Ranka Stanković, Miloš Utvić, Ivan Obradović and Božo Kolonja. “Managing Mining Project Documentation Using Human Language Technology”. *The Electronic Library* Vol. 36, no. 6 (2018): 993–1009. <https://doi.org/10.1108/EL-11-2017-0239>

- Утвић, Милош. “Анотација Корпуса савременог српског језика”. *Инфоотека* Vol. XII, no. 2 (2011): 39–51
- Utvić, Miloš. “Izgradnja referentnog korpusa savremenog srpskog jezika”. Phdthesis, Univerzitet u Beogradu, Filološki fakultet, Beograd, 2014, accessed August 1, 2019. <https://fedorabg.bg.ac.rs/fedora/get/o:10061/bdef:Content/download>
- Утвић, Милош. “Листе учестаности Корпуса савременог српског језика”. In *Научни састанак слависта у Вукове дане. Српски језик и његови ресурси: теорија, опис и примене. 3/43. научни састанак слависта у Вукове дане, Београд, 12-15. IX 2013.*, Милановић, А., Ж. Станојчић and Љ. Поповић, Vol. 43/3, 241–262. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2014
- Утвић, Милош В., Иван М. Обрадовић, Ранка М. Станковић, Александра Ђ. Томашевић and Биљана Ђ. Лазић. “Израдна специјалних корпуса савременог српског језика на примеру корпуса из области рударства”. In *Српски језик и његови ресурси: теорија, опис и примене. 3/47. научни састанак слависта у Вукове дане, Београд, 2017.*, Ђорић, Б. and А. Милановић, Vol. 47/3, 103–118. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2018. <https://doi.org/10.18485/msc.2018.47.3.ch7>
- Утвић, Милош В., Ранка М. Станковић, Александра Ђ. Томашевић, Михаило Ђ. Шкорић and Биљана Ђ. Лазић. “Претрага корпуса заснована на употреби екстерних лексичких ресурса путем веб-сервиса”. In *Српски језик и његови ресурси: теорија, опис и примене. 3/48. научни састанак слависта у Вукове дане, Београд, 2018.*, Ђорић, Б. and А. Милановић, Vol. 48/3, 279–298. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2019. <https://doi.org/10.18485/msc.2019.48.3.ch12>
- Vitas, Duško and Cvetana Krstev. “Processing of Corpora of Serbian Using Electronic Dictionaries”. *Prace Filologiczne* Vol. 63 (2012): 279–292
- Обрадовић, Иван, Александра Томашевић, Ранка Станковић and Биљана Лазић. “Увођење доменских и семантичких маркера за област рударства у српске електронске речнике”, In *Српски језик и његови ресурси: теорија, опис и примене. 3/46. научни састанак слависта у Вукове дане, Београд, 2016.*, Драгићевић, Р. and А. Милановић, Vol. 46/3, 147–158. Београд: МСЦ, Универзитет у Београду, Филолошки факултет, 2017. [http://doi.fil.bg.ac.rs/pdf/eb\\_ser/msc/2017-3/msc-2017-46-3-ch10.pdf](http://doi.fil.bg.ac.rs/pdf/eb_ser/msc/2017-3/msc-2017-46-3-ch10.pdf)