Digital Dictionary of the South Serbian Dialect

UDC 811.163.41'322 811.163.41'374'282.3

Miljana Mladenović
ml.miljana@gmail.com
University of Belgrade, Faculty of Mathematics

SUMMARY: Digital dictionary of the South Serbian dialect is the first comprehensive implementation of digital versions of a vocabulary dialect of the Serbian language, based on the dialect of southern Serbia, by Prof. Dr.Momčilo Zlatanović. It was published at www.vranje.co.rs and initially contained 10,950 terms. This is the first digital resource in the Serbian language, which, in addition to linguistic information, provides a number of others: sound information (pronunciation) of terms and examples of the use of words or phrases as they are spoken in the dialect; graphic information about the geographical location using concepts of Google Maps and Geocoding services; statistical information, which cannot be or is difficult to obtain using conventional vocabulary, of the etymological origin of the words (as it is from Turkish, Persian, Latin, etc.). It also contains information on certain types of words, use of words, origin, meaning, etc.; linking vocabulary and sharing its content through social networks. Another important aspect of the dictionary is that it works as a wiki resource, that is, it allows web users to expand and complement the dictionary in three ways: by adding entries that are not in the dictionary, introducing and applying examples of use of existing concepts and commenting and pointing to new meanings, creating new etymological templates and adding new toponymic characteristics related to the origin and the application range of the given terms. Tools were developed to provide general information on the trend of the growth and use of vocabulary - frequency of search terms, the current number of terms, the number of users working on the development of vocabulary and so on.

KEYWORDS: The Prizrensko-Timocki dialect, the Dialect of Vranje, Digital dictionary, Dialectal dictionary

DATE OF SUBMISSION: 26 February 2014
DATE OF ACCEPTANCE: 4 May 2014

1 Introduction

This paper introduces a digital dictionary varieties of Serbian language known in linguistics as the Prizrensko – Juznomoravski dialect.[1] It is, in fact, part of the Shtokavian Prizrensko-Timocki dialect of Serbian language prevalent in the southern regions of Serbia. The basis for the construction of a digital dictionary is the third edition of the Dictionary of southern Serbian Dialect (Zlatanović, 2011). The author of the paper version of the dictionary has been working on that edition and collecting of concepts and examples in the speech, for more than 10 years. Words that are found in the dictionary the author collected by himself

1 Dialect is a word of Greek origin (διάλεκτος, dialektos) and refers to the variety of a language used by particular groups of people belonging to a particular geographical area or social, professional, ethnic, or other category. Dialect has its own vocabulary that is more or less different from the base, and can have its own grammar and phonology

mostly working in the field and can be said to be in daily use in locations of: Vranje, Poljanica, Pčinja, Preševska Moravica, Preševska Crna Gora, Grdelička klisura, Vlasina, Crna Trava and some others. Geographically, it is the territory of the whole of Pčinja and a part of Toplica region. The dictionary also includes certain words taken from the "Travelogue of Hadži-Anta Kalimanac" (Hadži-Vasiljević, 1910) for which it was stated that it had been written based on "the right and faithful speech of Vranje". The author has noted down some of the entries thanks to the internal notes of professor of Russian language in high school "Bora Stanković" in Vranje, Tihomir Stefanović. They carry an additional label (T. S.) in the dictionary.

Unlike the Serbian literary language, the dialect of southern Serbia contains 32 sounds (phonemes). Letters to two additional sounds are choices of the author. One is a "soft voice" that originated from the Old Church Slavonic and it is tagged by (grapheme) **b**. The second is the digraph **dz** which has the grapheme **д̂з** in this dictionary. In a speech in southern Serbia certain words of foreign origin are also used. By searching through this dictionary it can be easily found that there are many words, primarily of Turkish origin, but also of Latin, German, Greek, Sanskrit etc.

2 Digital resources and tools for the study of dialects

Serbian language is based on Shtokavian diasystem (language system is derived based on the characteristics of a language in macro space) of Ekavian and Jekavian dialects (Okuka, 2008). Dialect is a language system that has a high degree of similarity with diasystem of one geographical area, and on the other hand has a system of its own characteristics that differentiate it from other dialects of the same language. A dialect may be constituted of subdialects in some micro space and of mini dialects (local speech). Each dialect is characterized by geographical area in which it is used and by linguistic features (phonological, morphological, lexical and syntactic).

As one of the dialects of Serbian language, the Prizrensko-Timocki dialect is used on the territory which to the Southwest has borders with Albania, to the south with Macedonia, to the east with Bulgaria, and to the north extending to Stalać. It is, because of internal differences (Ivić, 1985), divided into sub dialects: Prizrensko-Juznomoravski, Timocko-Luznicki and Svrljisko-Zaplanjski.

With the development of computer science, the study of dialects received a significant boost in the form of software tools and digital resources used. In addition to the development of digital dictionaries of dialects (Karanikolas et al., 2013; Keymeulen et al., 2013; Ćavar et al., 2000), it is showcased through the digitization of manuscripts vocabularies (Benacchio et al., 2012), the development of tools for the management of dialectal dictionaries (Pereira and Gillier, 2012) and the software for data visualization and linguistic dialectal maps - atlases (Sibler et al., 2012; Petsas, 2009). Modern geographical information systems (GIS) provide sophisticated and efficient spatial data analysis. However, for a long time, a field of linguistics was not the focus of research geographers. When linguists had begun to use GIS technology in the development of linguistic atlases, there has been an expansion of the branch of science that deals with the analysis of the geographical distribution and structure of language - geolinguistics. Analysis of geographical information, relevant for linguistic research, is related to the analysis and manipulation of spatial data, spatial statistical analysis and spatial modelling (O'Sullivan and Unwin, 2010).

The aim of this paper is linking of online GIS tools with the digital dictionary of the Prizrensko-Juznomoravski dialect in order to display the geographical prevalence of dialects and in order to show the geographical location of use of every word in the dictionary.

3 The process of digitizing of the dictionary

The first step in the digitization of the Dictionary of the southern Serbian Dialect (Zlatanović, 2011) is related to the preparations for the scan representing an iterative procedure to determine the optimal ratio of the output file size and quality of the resulting scan. In addition to the optimization of the parameters, it is important to perform a high resolution professional scanning. For this procedure FUJITSU Image Scanner fi-Series fi-5220C with an optical resolution up to 600 dpi was used. A total of 550 scanned pages in TIFF format was made. There were 32 folders made according to starting letters of terms in the dictionary and redistribution of files was performed.

Another important step was related to optical character recognition. For this purpose we used *Abby FineReader 11* (Abby FineReader, 2011). This software for optical character recognition techniques is based on machine learning. To ensure effective learning phase, we had to first identify the set of letters that will be used in the process of learning as well as in recognition. *Abby FineReader 11* provides recognition of 168 different natural languages (recognizing both of the Serbian alphabets: Cyrillic, Latin), four artificial (such as Esperanto and Interlingua language) and seven formal languages (C / C + +, Java, Pascal etc.). In the vocabulary there are alphabets: Cyrillic alphabet with accents (**пади́нче** - diminutive of a word *ūагина* (hillside), **надлиѣа́ње** - from the verb *nadletati* (fly above), **џалдиса́** - run away, run off, etc.), words and phrases written in Turkish alphabet (**Güzel** - primp, **perçem** - toupee, etc.) but also a number of words given in the Latin alphabet (usually these are the Latin names of plants and fungi, which are local names given by the dialect: **ви́лино кли́нче** - meadow mushroom Marasmius oreades). In addition to the languages which naturally use the Latin alphabet, in this dictionary there are certain words that come from Greek also written in Latin. For example: Gk. **Faétōn** - Exposed carriage on four wheels, Gk. **krommydi** - onion etc. Also, there are two graphemes mentioned in the introductory section: soft voice (**кла́дьнц** - water source, **бьбњеѣе́ње** - reverberation, etc...) and digraph dz (**д̂зь́вни** - echoes, **изд̂зьмбати** - to eat greedily etc.). It may be noted that all three alphabets contain diacritics. Based on the analysis of the use of letters in the dictionary, in the phase preceding the learning phase, simultaneously are activated the following alphabets: *Serbian Cyrillic*, *Serbian Latin* and *Turkish*. Given that the basic character set of the Cyrillic alphabet has no diacritics, accented characters are introduced in two ways: by expanding *Serbian Cyrillic* set for **Á É Ó Ô á é ó ô ý** and creating templates (patterns) for letters **и́ р́ ь́** and their corresponding *Normal*, *Bold* and *Italics* combination of uppercase and lowercase accented letters. At the end of the process a template sign for the digraph **д̂з** with reverse brevis has been created, which in the dialect of Vranje is the same as the letter ***s*** in the Macedonian language.

After the definition of a set of characters for learning, one could access the learning process. This stage is necessary to ensure a high level of recognition accuracy by observing the specifics of typography. It is known that the printing press brings some degree of distortion of characters so that they are never identical to their digital matrices. In the process of learning incorrect character recognition cases are recognized and eliminated. The process is iterative and the user is to decide on a sufficient number of iterations, as well about a learning character set. For the purpose of this project we have selected one page for each side of the first letter, and so formed a learning set containing a total of 32 pages.

In the next phase, i.e. in the optical text recognition phase, a single text document in Word format from a set of 550 *tiff* documents has been derived, where each document corresponds to a glossary. The accuracy of the recognized text is significantly improved compared to the results that we received prior to learning. However, two main groups of problems arising from the recognition process needed to be solved. For the first group of problems it had not been possible to be eliminated by the automatic procedure of error recognition, but the second group had been possible to be recognized automatically, and therefore removed. The first group of problems includes errors caused by:

1. Occurrence of excess (non-existent) characters caused by dirt on the scanned document. In Figure 1, an example marked by number 1 refers to this type of problem. The sign "backslash" \ appeared due to the existence of a stain in the original document. (It should be noted that this source of errors can be reduced by using cleaning techniques on *tiff* documents using some of the filters to remove noise: Reduce Noise, Median and Gaussian Blur in Adobe Photoshop.)
2. Incorrectly recognized letters, punctuation, and incorrect or unrecognized accentuation. In Figure 1, an example designated by reference numeral 2 refers to this type of problem. The term **кајсиче** was not recognized as accented. In the original document stands **кајси́че**.
3. Unsuccessful detection of the end of a paragraph (each entry must be in a new paragraph). In Figure 1, an example designated by reference numeral 3 refers to this type of problems, but also an example marked by reference numeral 2, points to the same error.

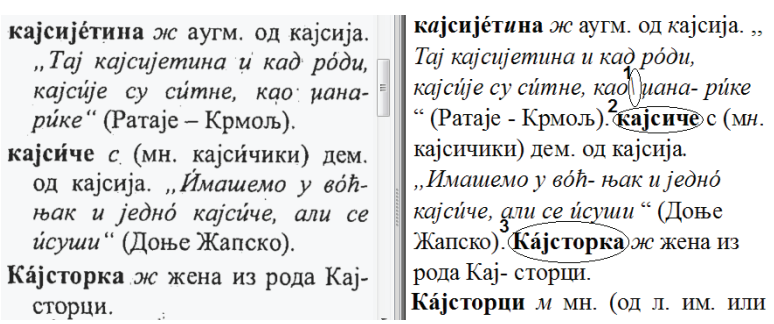
Sources of errors of the first group of problems are not uniform, not predictable and do not have a consistent logic of appearance. For this reason, they have been removed by hand. The second group of problems are:

1. Problem of hyphenation at the end of the line that have to be removed (In Figure 2, examples marked by number 1 are related to this type of problem, while the one marked by number 2 is not in this class).
2. Identifying numbers instead of letters, and vice versa (e.g. number **3** can be recognized instead of the letter **З** or the number **0** instead of the letter **О**).
3. Inadequate recognition of characters which are written in the same or similar in Cyrillic and Latin. Since the terms in the dictionary are written in Cyrillic, it is necessary that they should be recognized as Cyrillic. But, in examples that explain etymological origin of a term and are written in the original alphabet as in the case of Latin names, names in foreign languages (French, Turkish, etc.) it is necessary for text to be recognized in Latin.

The instances of inadequate recognition:

- Latin **k** and Cyrillic **к**,
- Latin **u** and Cyrillic italics ***и*** (и),
- Latin **y** and Cyrillic **у**,
- Latin italics ***m*** (m) and Cyrillic italics ***т*** (т), etc.
- For example, the correct recognition has to give:
- word **доли́на**, and not a word **долúна**,
- word ***Отиде*** (Отиде), and not a word **Отиде** (Отиде),
- word **дува́н**, and not a word **дувáн** ,
- herb "мајчина душица" **Thumus serpyllum L** – the Latin name should be written in Latin, and a letter **у** should not be recognized as a Cyrillic character.

Need for simultaneous translating of Latin and Cyrillic letters and their italics typographical representations cause a large number of incorrect recognition of characters of the same or similar form. However, in the case of vocabulary, high frequency of use of both letters did not allow the



кајсијéтина *ж* аугм. од кајсија. *„Тај кајсијетина и кад ро́ди, кајси́је су си́тне, као џана-ри́ке"* (Рataje – Крмољ).

кајси́че *с* (мн. кајси́чики) дем. од кајсија. *„И́машемо у во́ћ-њак и једно́ кајси́че, али се и́суши"* (Доње Жапско).

Ка́јсторка *ж* жена из рода Кај-сторци.

кајсијéтина *ж* аугм. од кајсија. „ *Тај кајсијетина и кад ро́ди, кајси́је су си́тне, као џана- ри́ке* " (Рataje - Крмољ). **кајсиче** с (мн. кајсичики) дем. од кајсија. *„Имашемо у во́ћ- њак и једно́ кајси́че, али се и́суши* " (Доње Жапско). **Ка́јсторка** *ж* жена из рода Кај- сторци.

Ка́јсторци *м* мн. (од л. им. или

FIGURE 1. The unpredictable sources of errors in the optical character recognition; left part shows the original *tiff* document to be recognized, and on the right is the text produced by recognizing process.

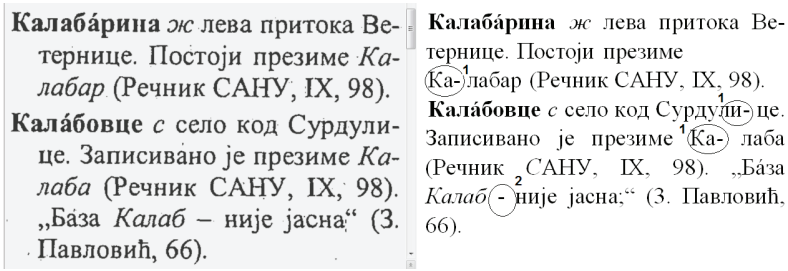


FIGURE 2: The problem of the existence of hyphenation at the end of the line; left part shows the original *tiff* document that has to be recognized, and on the right is the text produced by the recognizing process.

option of exclusion of any of them. This problem of the inability of even the visual inspection of application of different letters within a single word (e.g. in a Cyrillic word **дува́н** there is no way to visually determine whether **у** is a Latin letter Y) threatened to undermine the searching of the dictionary. If we would search for the word **дуван** in Cyrillic, and if it is in database, inscribed with the Latin alphabet Y, searches would not give results. The problem can be solved by the method of two-way transliteration.

For each word *word* in observed text *text* it is assumed that it is in Cyrillic. In the first step using the method *transliterateCtoL(word)* of the given word the same word written in Latin alphabet and appointed with *wordTrans* is generated. In the second step, taking the new word *wordTrans* and using the opposite transliteration method *transliterateLtoC (wordTrans)*, a word *wordTransBack* written in Cyrillic alphabet is generated. In the final step, we compare two words - *word* and *wordTransBack*, and if they are not equal, the word *word* in the text *text* is marked defective, i.e. marked yellow. The algorithm of bidirectional transliteration is given in pseudocode below. Table 1 shows examples of words for which it is presumed they are given in Cyrillic alphabet (first column), then they are converted by method *transliterateLtoC* to the Latin alphabet (second column), and then converted by method *transliterateLtoC* to their Cyrillic representations. In the cases of incorrectly written initial word (in first column), the contents of the first and third columns are not identical.

Algorithm 1: method of two-way transliteration

Input: The text of **n** unlabelled words for which it is desired to determine if it contains some words written by combination of the letters written in Cyrillic and Latin letters

Output: The text of **n** words, where there are **m** <= **n** labelled words that are written in a combination of Cyrillic and Latin letters

```
1.  for each (string word in text)
2.    string wordTrans=Empty;
3.    string wordTransBack=Empty;
4.    Boolean wordOK=true;
5.   // first transliteration step

6.    wordTrans= transliterateCtoL(word);
7.  // second transliteration step – transliterate back
8.  wordTransBack = transliterateLtoC(wordTrans);
9.    // if wordTransBack not equal to word, find mismatches
10.  if (word != wordTransBack){
11.    wordOK=false;
12.    Highlight(word, Color.Yellow);}
13. return text;
```

word	wordTrans	wordTransBack
долúна	**dolúna**	**долу́на**
дува́н	**dyván**	**д?ва́н**
hymus serpyllum	**х?мус серп?ллум**	**h?mus serp?llum**

TABLE1. Examples of detection of incorrect identification of Cyrillic and Latin letters, using the method of two-way transliteration

By the method of a two-way transliteration error recognition concerning Cyrillic and Latin letters is performed. A similar algorithm was designed for debugging recognition of digits as letters and vice versa, while the algorithm of recognizing words that were divided into syllables is based on the use of regular expressions in a form of .-\(S)* which can detect all instances of horizontal dashes "-" after any character. All three algorithms were implemented as a macro in *MS Word*. Figure 3 shows the appearance of the ribbon (part of the desktop *MS Word* that includes the icons of commonly used tools and options logically grouped on the basis of tasks of similar purposes) that is created for the work on the dictionary. Function *MarkiranjeSumnjivihZnakova* contains all three of the detection algorithms. The results of this macro-routine are shown in Figure 3 and are reflected in the different colored parts of the text. Green (number 1) mark the appearance of hyphenation, pink (labeled No. 2) indicates problems of recognizing digits and yellow problems of recognition of Cyrillic and Latin (numbered 3). Macro *Autokorekcija* automatically corrects all characters previously labelled by a macro *MarkiranjeSumnjivihZnakova*. Also there is a macro that removes all of the previously set labels in the text (in Figure 3 denoted by *ResetMarkiranja*). It should be noted that the appearance of normal and italic styles in a single word, if they are in the same alphabet are not considered faulty. The words in Figure 3 which are the combination of normal and italic styles in a single word but are cases of correct recognition are: меу́нчики, Врање, велике, мизики, etc.

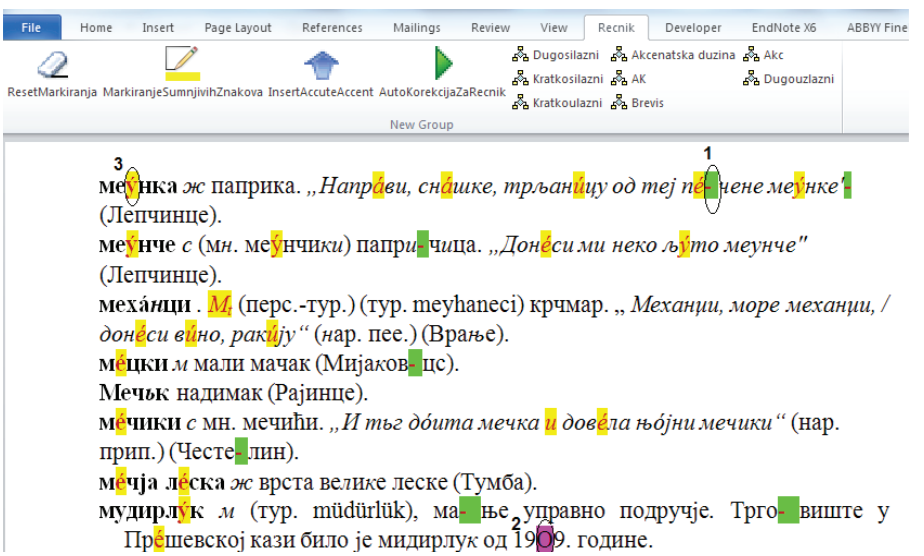


FIGURE 3. Macro-routines in *MS Word* for automatic detection and correction of problems concerning hyphenation at the end of the line and the two classes of errors in text recognition

4 Building a database of dictionaries

The database used in the implementation of the dictionary is a Microsoft SQL Server 2005. Preparing the loading of the base was carried out in two phases. In the first stage there is a text document, obtained by the process of optical character recognition, transformed into a *CSV* (*comma-separated values*) format which was used for creating database *Recnik.mdb* in *Microsoft Access 2010*. The reason of two phases of filling is based on the need for modelling and generating additional group of data in the dictionary, as well as for the extraction of certain types of data needed in generating of data tables of codes. For the purposes of the application you need the following data tables of codes: *Slovarnik*, *EtimološkoPoreklo*, *GeografskaLokacija*, *RodReči*. The role of codes in the application is to simplify the search function (they are used as data sources of drop-down list in the search forms) by methods of filtering database based on: first letter, the etymological origin of the word, geographical location of use

of the term and the concept of related words. For example, if you want to make a query to find all the words in the dictionary whose etymological origin is French, you have the option of selecting a filter from the list of all geographical locations used in the dictionary which gives the information that you wanted. Generating of data tables of codes can be done in two ways. The first method makes the application faster, but less flexible and is created by the extraction of data for each class of codes from the dictionary and their insertion in data tables. In the example above, the data table that contains data about etymological origin of words (*EtimološkoPoreklo*) would have data: *Albanian*, *Arabic*, *Greek*, *French*, *English*, etc. For each term which is added to the dictionary after creation of data tables of codes, for which there would be no corresponding data in some of the data tables of codes, there would have to be additional procedures to insert such kind of data. The second method is more flexible, but the system of inserting of the new terms into the dictionary is slower. In the realization of this vocabulary, we opted for the second way. Instead of a fixed data tables of codes, we just defined query phrases for creating all needed data tables of codes and for storing them in the internal memory (so-called "in Memory" data tables) when the application is activated. Created data tables are cached and remain unchanged in the application cache until the need for introducing of new information (e.g. a new datum of etymological origin) in one of the data table, when it is deleted from the cache and a new one is created. In the option we considered first, there is the problem of data redundancy, and in the other one, application occupies memory of the web-server to a higher degree.

In the second phase of the database filling, the database *Recnik.mdb* was exported into the production database *MS SQL Server*. After completion of the second phase, further operations on the database relate to the toughest part of this project: generating links to connect the database records to external files with audio content (subsection 3.2).

4.1 Metadata vocabulary

During the first phase of modelling of database, the dictionary was described by the following metadata:

- Term - a word or phrase in the dialect, written with an accent
- Description - a description of the meaning of the term which can include other information: part of speech, grammatical number, information about the derivation and inflection
- Gender- grammatical gender of the term
- Example - examples of the use of the term in the dialect speech
- Origin - the etymological origin of the term
- Location - the geographical location where the author noted an example use of the term
- Indication of cross-referencing
- Cross-reference (relationships among terms of the same or similar meaning.)

After completion of the second phase of the loading of the base, the dictionary has received six additional types of data:

- *The term written in Cyrillic without accents* (for the case when a user does not use diacritics in the formation of the query and does not use "soft voice" but the letter **a** as a substitution.)
- *The term written in Latin letters without accents* (used for generating the name of audio files and also used for search by a user request given in Latin alphabet. It has been obtained by using automatic transliteration of a term given in Cyrillic alphabet without accents into the term in Latin alphabet with no accent.)
- *Name of the audio file* with the purpose of generating an external link to a suitable audio track
- *The status* of the term in the dictionary
- *Username* of the author who inserted the term in the dictionary
- *Date of registration* of the term in the dictionary.

A program for generating the content of added types of data has been written: The term written in Cyrillic alphabet without accents (data column named *PojamCir*), the term written in Latin alphabet without accent (data column named *PojamLat*) and name the audio files (data column named *Audio*) are generated using the data in the data column named *Pojam* representing the term written in Cyrillic alphabet with accents. The other three fields are initially populated with the appropriate values. Figure 4 presents the look of the first 10 entries of the data table Dictionary, after the end of the process of filling the database.

S...	Pojam	Opis	PojamCir	PojamLat	Audio	Rod	V.	Vidi...	Primer	Poreklo	Lokacija	Status	A..	Datum
A	а́	за истицање (ве…	а.	а.	а		0		„Гла́ва, а у гла…		Собина	активан	MZ	NULL
A	а-а	узв. за исказива…	а-а	а-а	а-а		0		„Ина ли вода?" …		Мијовце	активан	MZ	30.06.2013 …
A	абы́д	коровска биљка…	абад	abad	abd	м	0		„Исе́чи сас ко́су…		Ратаје - Крмољ	активан	MZ	NULL
A	абьдљи́в	-а, -о који има а…	абадљив	abadljiv	abdljiv		0		„Таj ливáда је а…		Црна Река	активан	MZ	NULL
A	Абьдљи́вица	потес	Абадљивица	Abadljivica	Abdljivica	ж	0		„Мло́го ра́сне а…		Црна Река	активан	MZ	NULL
A	абьче	дем. од абьд.	абаче	abače	abce	с	0					активан	MZ	NULL
A	абаџи́ја	занатлија који и…	абаџија	abadžija	abadzija	м	0		„Милан је бија …	(ар.-тур.) (тур.…	Преображење	активан	MZ	NULL
A	абаџи́лък	"абаџијски зана…	абаџилак	abadžilak	abadzilk	м	0		„Дéда му се ба…	"(ар.-тур.) (тур…	Врање	активан	MZ	NULL
A	Абди́ница	њиве у Лопарди…	Абдиница	Abdinica	Abdinica	ж	0					активан	MZ	NULL
A	Абе́рка	надимак (Врање).	Аберка	Aberka	Aberka	ж	0			(ар.) (тур. habe…	Врање	активан	MZ	NULL

FIGURE 4. Data table Dictionary after populating the database

4.2 Generating audio recordings

Multimedia record is most commonly associated with the database by entering data on the relative or absolute URI address of the location of the server (or, in general, of the network) where such record is located. In data table Dictionary (Figure 4), data column named Audio presents data related to the relative paths or the names of multimedia files. Since the software for playing audio, *Adobe Shockwave Player* (Adobe, 2008), in its version of the plugin to web browser does not have support for Unicode, content of data column Audio cannot contain specific letters of the Latin alphabet with diacritics. The names of audio files are equal to corresponding items of data column Audio. For example, audio-file named *abadzija.mp3* that contains pronunciation of the term and examples of using of the word **абаџи́ја** have got the name according the value of *abadzija* contained in data column Audio.

For the purposes of this vocabulary audio file sized about 30 GB has been recorded, containing pronunciation of each term and with the correct pronunciation of the accompanying examples. Recording was entrusted to the theatre in Vranje, with the professor Zlatanović as a consultant. *Audacity 2.0.2-A Free Digital Audio Editor* software was used for this procedure. Work on the processing of audio recordings comprised computer correction (clearing of noise, change of tonal parameters) and segmenting of audio tracks and storing them into different audio files. We have controlled storage of audio recordings, using an auxiliary *MS Excel* document that pointed to errors in naming of audio files or in cases when an audio file for some reason was missing. Figure 5 shows the controlling form for validation of naming audio-recording and content of data column Audio which represents file names of audio-recordings.

	F	G	H
	Folder	Path ="E:\eRecnik\govor\" & F5 & "\" & A5 & ".mp3"	FileExist
	A	E:\eRecnik\govor\A\a-a.mp3	TRUE
ija	A	E:\eRecnik\govor\A\abadzija.mp3	TRUE
ilk	A	E:\eRecnik\govor\A\abadzilk.mp3	TRUE
	A	E:\eRecnik\govor\A\abcee.mp3	FALSE
	A	E:\eRecnik\govor\A\abd.mp3	TRUE
ica	A	E:\eRecnik\govor\A\Abdinica.mp3	TRUE
/	A	E:\eRecnik\govor\A\abdljiv.mp3	TRUE
vica	A	E:\eRecnik\govor\A\Abdljivica.mp3	TRUE

FIGURE 5. Control of naming of audio-recordings

5 Building of the application

Web application for the vocabulary management can be found at www.vranje.co.rs web address. The basic functionalities of an application are:

- Basic search (search by a term),
- Semantic search,
- Search by creating of complex logical queries,
- Listing of terms by first letter,
- Pronunciation of terms and examples of use of terms,
- Marking of search results on a geographical map,
- Statistical information about the dictionary.

Figure 6 presents the form of a basic search, the search results in text and graphical form. Basic search includes the possibility that the user types a word, a part of a word or phrase. He can specify whether the search will be carried out for the terms in dictionary which: start with typed word, contain typed word or equal to a typed word. Search results provide information on the number of terms found in the dictionary that satisfy the given condition. Information written below are details about each of the found terms. If a term contains information about the geographical location of the use of the term (which means that the author has observed the use of the term on specific geographical location), then that information transforms into software link to the point at geographical map so that it is possible to click to view a section of geographical map with marked the location. It is known that *Google Geocoding* web service is a tool for finding geographical coordinates (latitude and longitude) from other geographical data, such as place names, street addresses, or zip codes. When software finds the corresponding geographical coordinates of the geographical locations that have been shown (as shown in Figure 6 those are **Биљача** and **Врањска Бања**), he sends them to another web-service - *Google Maps*. *Google Maps* is a tool that uses digital maps and based on the given coordinates can mark the desired point and display it in a given shape on the geographical map.

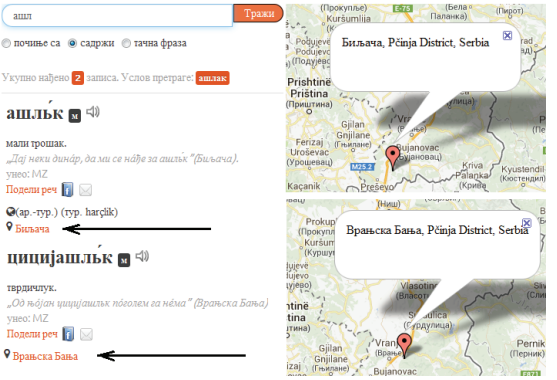


FIGURE 6. Results of a basic search

By implementing the techniques of research and extraction of text that rely on regular expressions, we created phrases for search that can perform vocabulary search based on semantic concepts. For example it can be required to find all examples of figurative speech, homonyms, place names etc. Figure 7 shows the different possibilities of semantic search through vocabulary. The current version of the application offers 11 groups of semantic search, and the picture shows the first set of results based on a given query: **Verbs = aorist**.

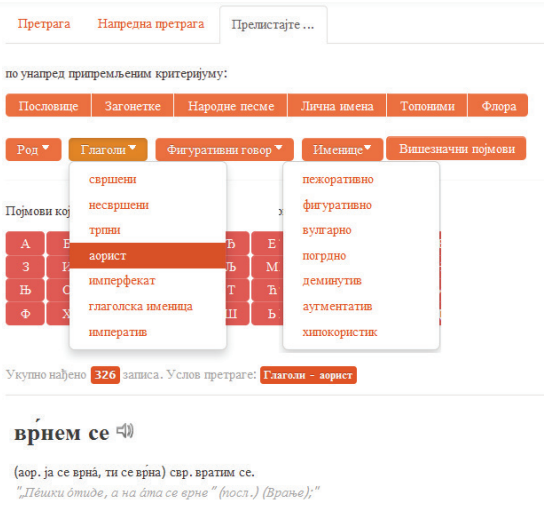


FIGURE 7. Semantic searching through the dictionary

Web users have the possibility to optimize the dictionary in any of the following three ways; by adding new terms into it; by expanding the vocabulary supplying it with additional examples of use and application; by making comments (making observations, highlighting certain aspects of meaning, showing new words and additional etymological characteristics). Access control is performed by restricting access and use to user groups and according to authorship (via user name).

User groups defined by this application are:

- anonymous Web users - have the ability to search
- authorized Web users - have the rights to add new terms into dictionary and make changes on terms which they already have added. Concepts that they have added have status "proposed".
- lexicographers - have rights to add new terms in dictionary, update and delete any terms. Also they can change the status of any of concepts from the "proposed" to "active" value, after which authorized users can not change such a term any more.
- administrators - manage system parameters and user groups.

In developing web applications we used the following software resources: relational database MS SQL Server 2005, ASP.NET Framework 4.0, C #, jQuery 1.8.9, AJAX, Twitter Bootstrap Framework 2.0.2., The Google Geocoding API, Google Maps API in .2, Social Plugins for Facebook, G+, Twitter, Digg, ShareThis Plugins, CSS3.

Web application includes the implementation of Web Services connecting to the currently most important social networks such as Facebook, Twitter, Google+, LinkedIn and Digg. It can be said that this is the first digital dictionary in Serbian language whose words and expressions you can exchange and share with friends on social networks and via email. Figure 8 showcases the implementation of a ShareThis service which is used for automatic providing of the term and its meaning if you want to share it online or by e-mail (in this case it is the phrase **аль́ш-верь́ш**).

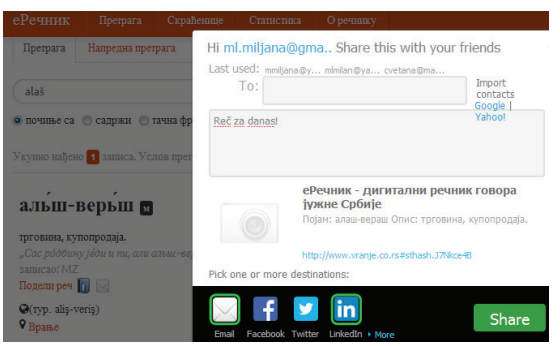


FIGURE 8. Sharing of vocabulary through social networks and email

6 Help in the study of characteristics of the Prizrensko-Juznomoravski dialect

The Digital Dictionary developed in this project is one of the sources for the study of nature of the dialect of Southern Serbia. Here we will mention a few specifics that can be reached by using different types of searching through vocabulary.

a) sequence of tenses by changing of accents
In the dialect, as opposed to the literary language, it is possible to build Past tenses without using auxiliary verbs, but changing verbs accents. It should be noted that without the accent, it is not possible to determine the tense of the action because the sentences are identical. For instance: a sentence in the dialect "**Зборѝ му, зборѝ, али он ништо не прѝфати**." (зборѝ – imperfect of a verb speak, прѝфати - aorist of a verb accept) is equivalent to the sentence in literary Serbian in Past tense "Говорио сам му, говорио, али он ништа није прихватио." (I had advised him, but he did not accept). But, the same sentence with different accents "**Збо́ри му, збо́ри, али он ништо не**

прифа́ти." is equivalent to a sentence in literary language in Present tense "Говорим му, говорим, али он ништа не прихвата." (I advise him, but he does not accept). The second example is a sentence in the dialect "**На моти́нке наређа́мо веша́љке и суши́мо за зиму.**" equivalent to a sentence in literary Serbian in Past tense "На мотке за сушење меса смо наређали каишеве меса и сушили смо их за зиму." (At the poles for drying of meat, we had piled belts of meat up and dried them for the winter). The same sentence in Present tense in the dialect is "**На моти́нке наре́ђамо веша́љке и су́шимо за зиму**", and in literary language is "На мотке за сушење меса наређамо каишеве меса и сушимо их за зиму." (At the poles for drying meat, we pile up meat belts and dry them for the winter).

b) verb anaphora – (**онодо́ујем, поонодо́ујем, прионодо́ујем, заонодо́ујем, заоно́дим**)
Verb, which is a replacement for each other verb only if they are in the same sentence, or if it is situated in the sentence after the one that contains the verb to which it relates. It can also be a substitute for any verb which is a topic shared between two interlocutors.

- An example of using of verb anaphora **онода́ја** in the sentence after the one in which the verb is in relation to the verb **Решава́ја**:

  in the dialect - **Решава́ја зада́ци цел дан ње́кња. Сви ги сам онода́ја.**

  in literary language - Пре неки дан, решавао је задатке целог дана. Све их је сам решио. (A few days ago, he worked on the math tasks all day long. He had solved them all by himself.)

- Example of use in a speech when respondents know about the action (verb).

  in the dialect –**„Пооно́ди гу још ма́лко ба́шчу."**

  The meaning of this sentence can be different, depending on what the recipient of the information had been previously working in the garden. The source of the sentence suggests that this work should be continued for a while. For example, it may be:
  „Заливај још мало башту." (Just keep watering a garden for a while) or
  „Окопавај још мало башту." (Dig a garden for a bit longer!) or
  „Огради још мало башту." (Build a garden fence some more!) etc.

c) Four degrees of comparison of adjectives
Apart from the basic three degrees, there is a fourth, which is located below the semantic positives, i.e. carries less expressiveness than the basic. For example, a scarf can be **шаре́на** - multicolored (positive), **по́шарена** (comparative), **на́јшарена** (superlative), but it can be **пришаре́на** - what is understood to be less than colorful. Similarly with the most descriptive adjectives, for example - beautiful:
прию́бав or **приубави́чав** - which means to be less than beautiful,
у́бав (beautiful),
по́убав (more beautiful),
на́јубав (the most beautiful).
There are other examples of described degree of comparison of adjectives: **прилу́т, прима́л, прила́дан, примале́чак, прису́тан, прискр́жав, прира́нке, прикисели́чав**.

d) Gradation of imperfective verbs
In Serbian literary language there are perfective verbs that are related to changing of states. They are formed by using of prefixes "**про**" or "**при**": **прогледао** (became able to see), **пропевао** (became able to sing), **приговорио** (complained). In the dialect of southern Serbia, there is a possibility of using the same prefixes in another way - to mean imperfective action with a small intensity:
те́пам несвр. тучем (imperfective - beating a someone)
протепу́јем несвр. тучем (imperfective - beating someone in a small intensity)

ле́ти несвр. лети (imperfective - flying)
проли́ћа несвр. лети (imperfective – lasting/appearing in a small intensity, usually for a snow or rain)
ту́рам несвр. стављам (imperfective - adding)
притуру́јем несвр. стављам, додајем (помало) (imperfective – adding bit by bit)

e) A larger number of consonants in continuity in a word
дзрдзо́љак – man of small stature and thin
дзрдзо́љче (pl. **дзрдзо́љчики**) – small and skinny child
Дзрдзинци – a family name in a village Bujkovac
Дзрдзцке – a name of a field in a village Dubnica
издзьмбати – eat greedily

Tools for the analysis of the vocabulary which are available to users differ depending on their privileges in the application. Unregistered user can use the advanced search section to get some statistics. For example, if a user wants to know how many words in the dictionary comes from a foreign language he can make query over metadata **Порекло** that applies the operation "contains" to one of the following values obtained by extraction from the dictionary (алб., ар., грч.,.) which means: Albanian, Arabian, Greek, etc. For example, if the selected value is „нем." (German) the query will be:

Порекло садржи: нем.

A query can be in a complex form of the logical **AND** expression. For example, if you would like to know how many words in the dictionary is feminine (**ж**) and also have German etymological origin (нем.), the query would be:

Порекло садржи: нем. Род једнако: ж

The obtained results are presented numerically and show the number of found words as well as a list of those words.

Query that can give all the adjectives that have the prefix, which we described in item c) of this section is:

Појам садржи: при Опис садржи: -а, -о

Those users who have privileges of lexicographer can browse log-files from which they can obtain concepts or groups of concepts that Web users are typically looking for. A set of records in log-files is provided in the form of:

datum:19/07/2013 vreme:16:17 pojam: đukavac
datum:19/07/2013 vreme:16:17 pojam: vjuge
datum:19/07/2013 vreme:16:18 izvor: Proverbs
datum:19/07/2013 vreme:16:19 izvor: palamar

From the log-files data structure analysis of frequency and types of searches over the vocabulary can be performed.

For all users statistic data on the number of personal names in the dictionary are available, the number of words of which the sources are folk songs, the number of words of which the sources are proverbs, the number of words which originate from the puzzles, the number of terms that represent figurative language use and so on. Also the geographical Google map is available, with the highlighted area in which the dialect of southern Serbia is used.

7. Conclusion

The primary goals of the Web application that was described in Section 5 are:

- building of significant digital resource of the dialect of southern Serbia, accessible to all internet users in a new and inventive way
- stimulating of the Internet community in the process of accruing, augmenting and preserving the dialect of southern Serbia
- creating of digital recordings of a speech in the dialect
- integrating vocabulary with geolocation resources (Geocoding API, Google Maps API v.2)
- integrating of vocabulary with social networks
- availing content sharing of vocabulary electronically
- inspiring digitization of dictionaries

of other dialects: "Нишлијски говор и речник", "Црнотравски говор", "Речник пиротског говора", "Народни говор и речи из власотиначког краја" (dialects of Niš, Crna Trava, Pirot, Vlasotince).

The application has been designed and implemented fully so that it is possible to expand the existing online dictionary and to add new vocabularies thereby expanding digital database of Serbian dialects. Further work on this dictionary will develop along two lines: to increase the number of terms, examples and sound files, and to enhance the functionality of existing application and develop it for smartphones and tablets.

Acknowledgements

Realization of eRečnik was whole-heartedly supported by Register of National Internet Domain Names of Serbia (RNIDS) in the project (4PI) in 2013. The authors of this project owe gratitude to actresses of the Theatre "Bora Stanković" in Vranje, Radmila Đorđević and Milena Stošić.

References

Abby Fine Reader: Version 11 Users's Guide. CA: ABBYY Software Ltd, 2011.

Adobe: Adobe Director 11 User Guide. CA: Adobe Systems Incorporated, 2008.

Benacchio, Rosanna, Han Steenwijk, Željko Jozić and Nada Vajs Vinja. "Digitalna obradba rukopisnoga rječnika Vocabolario di tre nobilissimi linguaggi, italiano, illirico, e latino Ivana Tanzlinghera Zanottija (1651.—1732.)". *Filologija* No. 58 (2012): 19–38.

Ćavar, Damir, Alexander Geyken and Gerald Neumann. "Digital Dictionary of the 20th Century German Language", In *Jezikoslovne Tehnologije za Slovenski Jezik: Proceedings of JS*, eds. T. Erjavec and J. Gros, 110-114. Ljubljana: Institut Jožef Stefan, 2000.

Ivić, Pavle. *Dijalektologija srpsko-hrvatskog jezika*. Novi Sad: Matica srpska, 1985.

Хаџи-Васиљевић, Јован. *Путопис Хаџи-Анте Калиманца*, Београд, 1910.

Karanikolas, Nikitas N., Eleni Galiotou, George J. Xydopoulos, Angela Ralli, Konstantinos Athanasakos and George Koronakis. "Structuring a Multimedia Tri-Dialectal Dictionary". *In Lecture Notes in Computer Science*, vol. 8082, 509–518. Berlin: Springer, 2013.

Okuka, Miloš. *Srpski dijalekti*. Zagreb: Prosvjeta, 2008.

O'Sullivan, David and David J. Unwin. *Geographic Information Analysis*. New Jersey: John Wiley & Sons Inc., 2010.

Pereira, Sandra and Raïssa Gillier. "TEDIPOR: Thesaurus of dialectal Portuguese". *In Proceedings of the 15th EURALEX International Congress*, 267-281. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 2012.

Sibler, Pius, Robert Weibel, Elvira Glaser and Gabriela Bart. "Cartographic Visualization in Support of Dialectology". In *The 2012 AutoCarto International Symposium on Automated Cartography, Columbus, Ohio, USA, 16 September 2012 - 18 September 2012*. Ohio: Cartography and Geographic Information Society, 2012.

Spyridon, Petsas. "Visualising Perceptual Linguistic Data". MSc in GIS dissertation, University of Edinburgh, UK, 2012.

Van Keymeulen, Jacques, and Veronique De Tier. "The Woordenbank Van De Nederlandse Dialecten." In *3rd eLex Conference. Electronic Lexicography in the 21st Century: Thinking Outside the Paper, Proceedings*, ed. Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langements, and Maria Tuulik, 261–279. Ljubljana: Trojina, Institute for Applied Slovene Studies, 2013.

Златановић, Момчило. *Речник говора југа Србије*. Врање: Аурора, 2011.

Web Addresses

Audacity. http://audacity.sourceforge.net/onlinehelp-1.2/reference.html
Shockwave. http://get.adobe.com/shockwave/
Geocoding API. https://developers.google.com/maps/documentation/geocoding/
RNIDS. http://www.rnids.rs
4PI. http://4pi.rs/