A WordNet Ontology in Improving Searches of Digital Dialect Dictionary

Miljana Mladenović, Ranka Stanković, Cvetana Krstev

A WordNet Ontology in Improving Searches of Digital Dialect Dictionary | Miljana Mladenović, Ranka Stanković, Cvetana Krstev | New Trends in Databases and Information Systems: ADBIS 2017 Short Papers and Workshops - SW4CH (Semantic Web for Cultural Heritage) | 2017 | 767 |

10.1007/978-3-319-67162-8_37

http://dr.rgf.bg.ac.rs/s/repo/item/0001871

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

Miljana Mladenović[1], Ranka Stanković[2], and Cvetana Krstev[3]

[1] eVox Solutions, Belgrade, Serbia,
ml.miljana@gmail.com,
[2] University of Belgrade, Faculty of Mining and Geology,
Djusina 7, Belgrade, Serbia
ranka@rgf.bg.ac.rs
[3] University of Belgrade, Faculty of Philology,
Studentski trg 3, Belgrade, Serbia
cvetana@matf.bg.ac.rs

Abstract. In this paper we present a method for automatic generation of a digital resource which connects all indirect synonyms of a dialect term to all indirect synonyms of a corresponding term in the standard language aiming to improve search over a digital dialect dictionary. The method uses SWRL rules defined in the Serbian WordNet ontology to identify sets of synonymous words. It also uses e-dictionaries to produce correct lemmas in the standard language that users usually use for search. The method was applied and evaluated on verbs and a group of nouns derived from verbs (verbal nouns). We compared results obtained by the system with those produced by humans and achieved the accuracy of 89.7%.

Keywords: dialect dictionary, ontology, WordNet, e-dictionary, South Serbian dialect

1 Introduction

The study of dialects has received a new impulse with the development of many software tools and digital resources for maintaining, enhancement, sharing, visualization and analysis of digital dialect dictionaries. This includes: the development of digital dictionaries of dialects [4], [13], [1], the development of tools for the management of dialectal dictionaries [10], the software for data visualization and presenting of linguistic dialectal maps [12], [11], the analysis of the geographical distribution of a language and geographical information relevant for linguistic research [9], using of Semantic Web-based techniques for representing digital resources as knowledge based resources and as Linked Open Data (LOD) on the Web [6], [15].

Digital dictionary of the South Serbian dialect[4], containing over 20 thousand terms, is the first comprehensive implementation [7] of a digital version of a dialect vocabulary of the Serbian language, produced on the basis of traditional

[4] On-line at http://www.vranje.co.rs

dialect dictionaries [16],[17]. This is the first digital resource for Serbian which, in addition to linguistic information, provides also: sound information (pronunciation) about 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; the etymological origin of the words, morphological information like part of speech, and additional semantic data. The content of the dictionary can be shared through social networks. Another important aspect of the dictionary is that it allows Web users to expand and complementit. Tools were developed to enable search of terms in three ways: (a) search by a term, (b) search by logical queries created over metadata, (c) terms browsing by the first letter. Search by a term implies that the user types a word or its part in the dialect. He/she can specify whether the search will be carried out for the terms in the dictionary which: start with typed word, contain it or are equal to it. Search results offer information on the number of terms found in the dictionary that satisfy the given query. This kind of search is standard for on-line dictionary look-up, but it is based on the presumption that a user knows what she/he is looking for which need not be the case for user not familiar with a dialect. This problem often encountered by students of a foreign language can be solved by explaining terms not known in a foreign language by expressing the same concepts in a language they are familiar with. Two other ways of search (search by creating a logical query over metadata and browsing of terms by a first letter) are more convenient to a novice, but they produce much more information than is expected or needed, which slows down the learning process.

In this paper we propose a method for connecting the standard language and the dialect, that would enable search over a digital dialect dictionary by using terms in the standard language. In Section 2 we discuss some previous approaches to searching digital dialect dictionaries. In Section 3 we represent resources used to improve searching performances of the digital dialect dictionary: Serbian morphological e-dictionaries used to produce all inflected forms of standard terms and Serbian WordNet (SWN) ontology represented in OWL2 format for which we define rules expressed in Semantic Web Rule Language (SWRL) to be used to generate synonymous groups in the SWN ontology on the basis of the indirect synonymy relation. In Section 4 we propose a method for automatic generation of a digital resource which connects all indirect synonyms of a dialect term to all indirect synonyms of a corresponding term in the standard language. In Section 5 the method is evaluated on verbs and a group of nouns derived from verbs – verbal nouns. Finally, we give some conclusions and directives for future work in Section 6.

2 The management of digital dialect dictionaries

The challenging task of digitizing a dialect dictionary can be solved in different ways considering software platform, database storage, search framework and additional management tools. It can be a Web application that uses relational databases, like Oracle used for storing the southern Dutch dialects [13] and

Joseph Wright's "English Dialect Dictionary" (EED) [8], MS SQLServer for storing a South Serbian dialect dictionary data [7], or it can be one of a machine-readable and interoperable Semantic Web standards, such as RDF, SKOS and SKOS-XL which are used, for example, in the case of the dialect dictionary of the German language [15] and two Austrian dialectal dictionaries [2]. When it comes to search techniques over a dialect dictionary, the metadata are usually used to search for specific dialectal information. For example, the retrieval of information in EDD can be limited to the structural units in the entries (heads, definitions, citations, comments, variants, etc.), while logical filters combining basic Boolean operators and metadata can be used for advanced search options. Similar techniques are used for retrieving information from the South Serbian dialect dictionary data [7]. Some digital dialect dictionaries like Dutch [15] have records in their database that contain, besides *original headword*, additional fields – *dutchfield headword* and *search term in standard Dutch* – for advanced search options. In this paper we propose a method which enables search not only with one term in the standard language, but with a set of synonym terms in order to improve search.

3 Resources

3.1 Use of morphological e-dictionaries

The first problem with search of verbs in dialect dictionary is the grammatical form of the headword of the lexical entry. Namely, grammatical form of the headword of the verb lexical entry in the dialect dictionary is the present tense first person singular, while user's intent is to search for verbs using their infinitive forms, these being headwords for standard language dictionaries. To support that kind of search, it was necessary to add an infinitive form, that is, to lemmatize both a dialect verb and verbs in the standard Serbian that were retrieved from its definition. For lemmatization task we used Serbian morphological electronic dictionaries and grammars developed within the University of Belgrade Human Language Technology Group [14].

Morphological electronic dictionaries of Serbian for NLP are being developed for many years now. In the dictionary of lemmas (DELAS) each lemma is described in full detail so that the dictionary of forms containing all necessary grammatical information (DELAF) can be generated from it, and subsequently used in various NLP tasks. Serbian e-dictionaries of simple forms have reached a considerable size: they have more than 140,000 lemmas generating more than 5 million forms and 18,000 multi-word lemmas [5]. Dictionaries contain mostly standard language but also some dialect lemmas.

An e-dictionary of forms consists of a list of entries supplied with their lemmas, morphosyntactic, semantic and other information, so it is possible to attach lemma for all inflected forms in the dialect dictionary that match a form in morphological e-dictionary. After separation of all synonyms aligned with a dialect lexeme (from the standard language or dialect), infinitive forms were attached to the original form.

Among 4,152 filtered entries having a dialect form followed by the list of words or phrases in standard Serbian, 3,452 entries were verbs and others were verbal nouns (gerund). For 3,452 verb entries 7,353 synonyms were detected – related words or phrases in the standard Serbian that describe dialect forms and that have the same meaning as corresponding dialect words. A few verb entry examples are:

- batalim | batalen ; ostavim ; napustim *to quit*
- batisujem | kvarim ; upropašćujem *to ruin*
- bednim se | lepo se odevam ; doterujem se to dress up
- begam | begaj ; ja bega ; ti bega ; begajeći ; bežim *to flee*

After lemmatization we obtained the following result:[5]

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

Lemma was assigned for 505 dialect forms out of 3,452 dialect forms given in the first person singular, present tense. Infinitive forms were assigned to 4,384 word forms in standard Serbian that were connected to dialect forms (out of 7,353). Word forms that were not lemmatized consisted of either of another dialect form e.g "ja znaja; ti znaja; znam_znati *to know*" not presented in e-dictioanries, or adjectives used to describe verbs e.g. "zgugurija se; zguguren; pogurim_poguriti se *to stoop, to be stooped*".

In the dialect dictionary the relation between verbal nouns and verbs was established in some entries: "šljakanje | od šljakati *to slap*", but it was not done systematically. Again we used morphological e-dictionaries in which all verbal nouns are marked with a special marker which enabled us to establish missing connections between verbs and verbal nouns. In that way 700 relation were established.

3.2 Calculating the set of near synonyms by using the WordNet ontology

The development of the lexico-semantic resource Serbian WordNet (SWN), is based on the semantic network Princeton WordNet (PWN) [3]. Today, SWN is a set of more than 22,000 concepts called synsets where a concept is represented by the set of synonymous word forms that have the same or similar meaning in a given context. Synsets respect the syntactic categories noun, verb, adjective, and adverb and can be interconnected by semantic relations, while word forms can be connected by lexical relations. In SWN ontology there are currently 2,243 verb synsets defined as ontology individuals belonging to the VerbSynset class:

```
<rdf:type  rdf:resource="&swn30;VerbSynset"/>
```

[5] Lemmatization was done using Unitex, the corpus processing system (`http://unitexgramlab.org/`)

We wish to defined rules that can be used to generate synonymous pairs of verbs found in the SWN ontology that were not be based only on the relation of direct synonymy. By doing that, we created a broader set of synonyms for each verb defined in SWN ontology. Relations that participate in finding the broader set of indirect synonyms are: *synonym, similar_to, also_see, verb_group, hyponym.* Total number of synsets relations in SWN: *synonym* 22,162, *similar_to* 371, *also_see* 242, *verb_group* 191, *hyponym* 21,554. Figure 1 shows an example of synsets and relations between them in the SWN ontology which are used to define a set of indirect synonymous concepts of the verb "dopustiti – permit, allow, let, countenance" by using hyponymy related verbs, synonyms of hyponymy related verbs, verbs belonging to the same semantic group of verbs as the observed verb, and by verbs defined as semantically similar to the observed one.

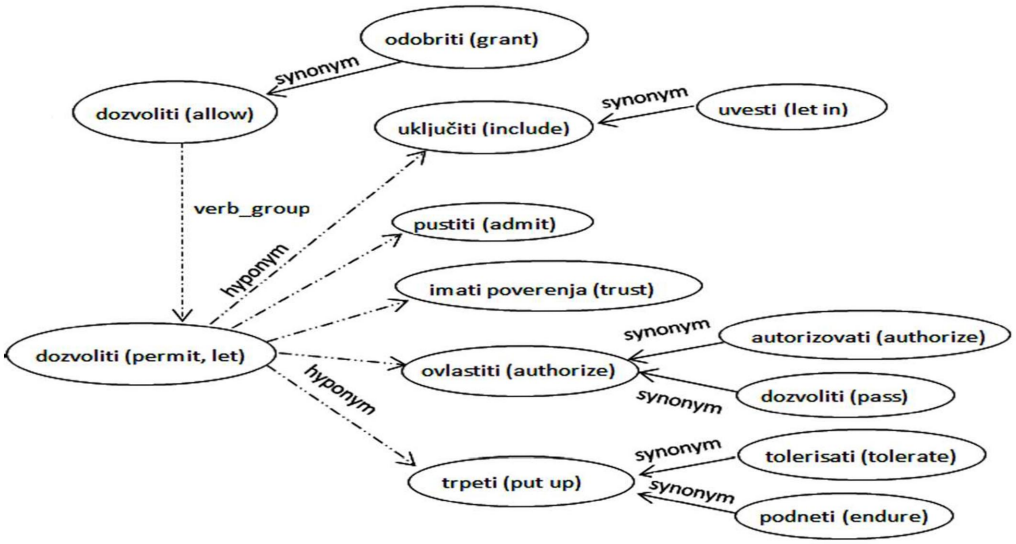

Fig. 1. Verb synsets in the SWN ontology which are mutually connected with relations that participate in finding a set of indirect synonyms of the verb "dopustiti - permit, allow, let, countenance".

Reasoning rules in the SWN ontology for determining the existence of indirect synonymous pairs in which the five previously mentioned relations participate and which can be used for generating the "indirectSynonymy" relation are presented here in the form of Jena rules:[6]

```
"[rule1:(?a eg:label ?b)(?a eg:synonym ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]"
"[rule2:(?a eg:label ?b)(?a eg:similar_to ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]"
```

[6] We used the following software tools in this paper: Developing tool Eclipse Java EE IDE Luna and Apache Jena open source software development environment which allows for reasoning at the level of OWL 2 language by converting OWL rules into the Jena rules format.

```
"[rule3:(?a eg:label ?b)(?a eg:also_see ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]"
"[rule4:(?a eg:label ?b)(?a eg:verb_group ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]"
"[rule5:(?a eg:label ?b)(?a eg:hyponym ?c)(?c eg:label ?e) ->
(?b eg:indirectSynonymy ?e)]"
```

By looking at the rules from the set {rule2,...,rule5}, it can be noticed that each of them can be expanded with the synonymy relation, yielding the following expanded set of rules:

```
"[rule6:(?a eg:similar_to ?c)(?a eg:label ?b)(?c eg:synonym ?d)
(?d eg:label ?e) -> (?b eg:indirectSynonymy ?e)]"
"[rule7:(?a eg:also_see ?c)(?a eg:label ?b)(?c eg:synonym ?d)
(?d eg:label ?e) -> (?b eg:indirectSynonymy ?e)]"
"[rule8:(?a eg:verb_group ?c)(?a eg:label ?b)(?c eg:synonym ?d)
(?d eg:label ?e) -> (?b eg:indirectSynonymy ?e)]"
"[rule9:(?a eg:hyponym ?c)(?a eg:label ?b)(?c eg:synonym ?d)
(?d eg:label ?e) -> (?b eg:indirectSynonymy ?e)]";
```

We have carried out several experiments using different lengths of chains of relations taken from the set of five given relations *synonym, similar_to, also_see, verb_group, hyponym* which enabled us to conclude that the sufficient length of a chain is 3. In that way we manually defined 24 rules (combinations of 4 relations taken 3 at a time) having the form illustrated by the following examples in which a chain is formed of relations *similar_to, also_see* and *synonym*.

```
"[rule10:(?a eg:similar_to ?c)(?a eg:label ?b)(?c eg:also_see ?d)
(?d eg:synonym ?e)(?e eg:label ?f)->(?b eg:indirectSynonymy ?f)]"
"[rule11:(?a eg:verb_group ?c)(?a eg:label ?b)(?c eg:also_see ?d)
(?d eg:synonym ?e)(?e eg:label ?f)->(?b eg:indirectSynonymy ?f)]"
"[rule12:(?a eg:hyponym ?c)(?a eg:label ?b)(?c eg:also_see ?d)
(?d eg:synonym ?e)(?e eg:label ?f)->(?b eg:indirectSynonymy ?f)]";
```

Restrictions that we introduces were the following: (1) a relation from the set of five given relations *synonym, similar_to, also_see, verb_group, hyponym* cannot be repeated more than once in a given rule; (2) the relation *synonym* has to be found only as the last one in a sequence of given five relations. In this way we obtained 33 rules for reasoning about the existence of the *indirectSynonymy* relation and, after inferencing, 6,430 indirectSynonymy related pairs of verbs.

4 The implementation of new search features

The proposed method for connecting the standard language with the dialect dictionary relies on a table of sets of synonymous words of the standard language which are related to an equivalent set of dialect entries. This table is used as a part of the Web tool for advanced search of the digital dictionary of the South Serbian dialect. Figure 2 shows resources and procedures included in the process

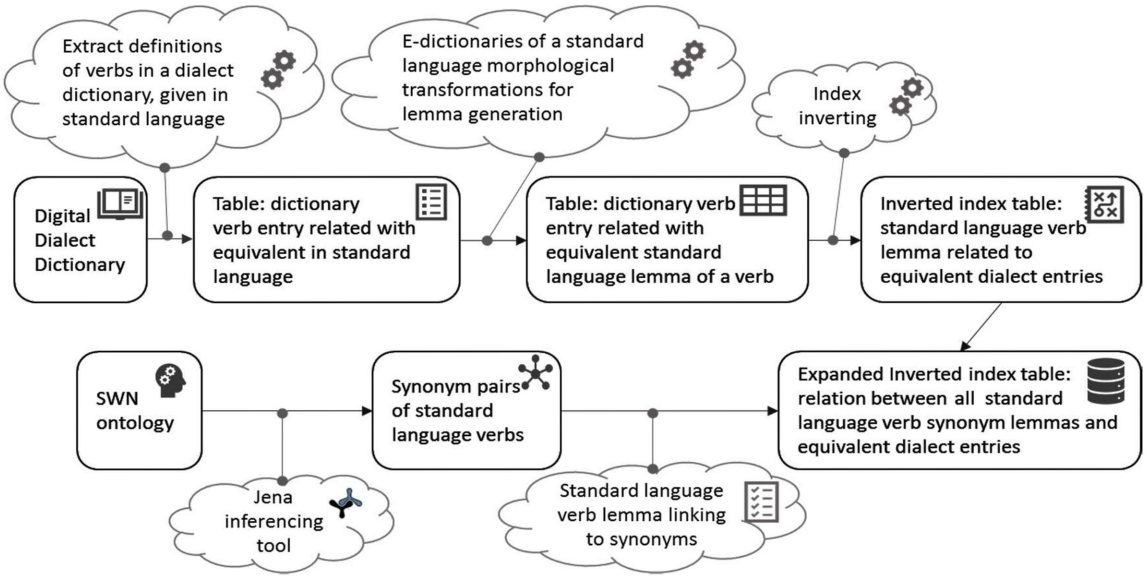

Fig. 2. Architecture of the system for building a resource that improves the dialect dictionary search tool.

of generating such a table named *Expanded inverted index table*. The table was created in five steps.
(1) Automatic extraction of definitions of verbs and verbal nouns in the digital dictionary was performed by Transact-SQL stored procedures since dialect dictionary storing database is MS SqlServer. Total of 4,153 entries representing verbs, verbal nouns and their definitions were extracted, which represents about 20% of the size of the whole dictionary. Two examples of a table rows (containing an entry and a definition) obtained in the first step are:

isabim "(imp. isabi; aor. ja isabi, ti isabi; r.pr. isabija, -ila, -ilo) svr. iskvarim, upropastim." *to ruin, to destroy*
ačkam "(imp. ačkaj; impf. ačkašem) nesvr. kotrljam." *to roll*

(2) E-dictionaries of the standard language were used for detecting verbs occurring in each definition and for transforming them from a form used in this dialect dictionary (the nominative case, singular, present tense, for example a verb *upropastim*) into the form used in contemporary dictionaries (the infinitive, for example *upropastiti*). Two corresponding examples obtained after the second step are (the infinitives generated by the transformation process are given in bold):

isabim | isabi; ja isabi; ti isabi; isabija; iskvarim_**iskvariti**; upropastim_**upropastiti**
ačkam | ačkaj; ačkašem; kotrljam_**kotrljati**

(3) An index table was created by inverting the table obtained in the step (2) i.e. all dialect dictionary entries related to each infinitive representing a verb in a standard language were found. In the case of verbal nouns, they were joined to verbs from which they were derived, and we subsequently treated in the same

manner as verbs. In the subsection 3.1 we noted that total of 4,152 headwords of the dialect were related to 4,384 infinitives of verbs in the standard language, of which 4,252 infinitives are unique. For that reason, this is the length of the inverted table (total number of records). One of the records is shown below. It is related to the first example in the step (2) and represents an infinitive of a verb *upropastiti* linked to the 8 entries found in the dialect dictionary whose definitions contained this verb in the form of the first person singular*upropastim*.

upropastiti | isabim batišem dokrajišem istrovim izabim izakam oznobim profućkam

(4) In this step the SWN ontology was used in order to implement inference rules defined in Subsection 3.2. As a result, the table was obtained that contained in each row a verb lemma and its indirect synonyms (in the standard language) as shown in the next example:

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

(5) In the final step, tables obtained in steps (3) and (4) were joined aligning an infinitive of a verb in the standard language representing a set of synonyms in the dialect by a verb in the standard language representing a set of synonyms obtained from SWN ontology. Next example represents two joined sets of synonyms aligned by the verb *upropastiti*.

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

On the left side of the vertical line are synonymous words in the standard dictionary while on the right side are synonym words in the dialect. The use of this table by the dialect dictionary advanced search tool enables a user to type any of eight words of the standard language in order to obtain all equivalent synonymous words in the dialect.

5 Evaluation

The evaluation of the proposed method can be observed as the evaluation of a classification task, so we have performed an estimation of the accuracy of pairing the digital dictionary entries with the standard language entries comparing results obtained by the system with results given by humans. Two language experts annotated the inverted table described in the step 3 of the section 3. The table was divided into two equal parts and each annotator marked one of them. They used 3 marks for each record in the table: 1 – if an infinitive of the standard language has the same or similar meaning as verbs of the dialect in the same record; 2 – if it is was not clear whether an infinitive of the standard language has the same or similar meaning as verbs of the dialect in the same record; 3 – if an infinitive of the standard language has not the same or similar meaning as

verbs of the dialect in the same record. At the same time, we made an automatic procedure to check if there were headwords of a dialect dictionary that were not related to any infinitive after inverting the table. This procedure classifies infinitives on those who take a part in relations (related) and those that do not (unrelated). When we compared human marks 1 with related, we obtained true positives. Human marks 2 and 3 compared to related gave false positives. In the similar way, comparing with the unrelated set produced false and true negatives. The confusin matrix is given in Table 1. Based on the confusion matrix, performance measures were calculated: precision $P = tp/(tp + fp) = 1.000$, recall $R = tp/(tp + fn)) = 0.874$, $F1 = 2PR/(P + R) = 0.933$, accuracy= 0.897.

Table 1. The confusion matrix of the process deciding whether dictionary entries are correctly aligned with standard language entries.

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

We can notice that the proposed method is completely precise, but the problem with high value of false negatives lies in the shortcomings that were found in the dialect dictionary. The most frequent are: writting erros, the lack of a verb in the definition, a verb is not written in a standard format (first case, singular, present tense), a verbal noun can not be linked to any verb. Also, some standard verbs from definitions were missing from e-dictionaries and some dialect verbs were misineterpreted by e-dictionaries.

6 Conclusion

In this paper, we propose a method for improving search of the dialect digital dictionary by offering the possibility to search with terms in the standard language. The method uses SWRL rules defined in the ontology based on the semantic network Serbian WordNet to identify sets of synonymous words for each verb and verbal noun defined in the ontology. The method also uses e-dictionaries of the standard language to extract word forms defining verbs in the dialect dictionary and to transform them into lemmas (used for search). The method generates a table joining two sets of synonym words – one originated from the dialect dictionary, another from e-dictionaries of the standard language – for each verb extracted from a dialect dictionary. The evaluation of the method, treated as a classification, compared results obtained from the system with data provided by humans. The accuracy measure was acc=89.7%. In future work, we will experiment with other parts of speech and we will try to expand the set of ontological rules used in this system.

References

1. Čavar, D., Geyken, A., Neumann, G.: Digital dictionary of the 20th century German language. Jezikoslovne Tehnologije za Slovenski Jezik: Proceedings of JS pp. 110–114 (2000)
2. Declerck, T., Wandl-Vogt, E.: How to semantically relate dialectal Dictionaries in the Linked Data Framework. In: Proc. of the The 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014). pp. 9–12. ACL, Gothenburg, Sweden (April 2014)
3. Fellbaum, C.: WordNet: an eletronic lexical database. The Mit Press, Cambridge, Massachusetts, EUA (1998)
4. Karanikolas, N.N., Galiotou, E., Xydopoulos, G.J., Ralli, A., Athanasakos, K., Koronakis, G.: Structuring a multimedia tri-dialectal dictionary. In: LNCS, International Conference on Text, Speech and Dialogue. vol. 8082, pp. 509–518. Springer Berlin Heidelberg (2013)
5. Krstev, C., Vitas, D., Stanković, R.: A Lexical Approach to Acronyms and their Definitions. In: Mariani, Z.V..J. (ed.) Proc. of 7th Language & Technology Conference. pp. 219–223. Fundacja Uniwersytetu im. A. Mickiewicza, Poznań, Poznań, Poland (November 2015)
6. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation 46(4), 701–719 (2012)
7. Mladenović, M.: Digital dictionary of the South Serbian Dialect. Infotheca 15(1), 42–55 (September 2014)
8. Onysko, A., Markus, M., Heuberger, R.: Joseph Wright's 'English Dialect Dictionary' in electronic form: A critical discussion of selected lexicographic parameters and query options. Language and Computers 69(1), 201–219 (2009)
9. O'Sullivan, D., Unwin, D.: Geographic Information Analysis. John Wiley & Sons Inc., New Jersey (2010)
10. Pereira, S., Gillier, R.: TEDIPOR: Thesaurus of dialectal Portuguese. In: Proc. of the 15th EURALEX International Congress. pp. 267–281. Norway (2012)
11. Petsas, S.: Visualising Perceptual Linguistic Data. University of Edinburgh, UK. (2009)
12. Sibler, P., Weibel, R., Glaser, E., Bart, G.: Cartographic Visualization in Support of Dialectology in Support of Dialectology. In: Proc. AutoCarto 2012: the international symposium on Automated Cartography. Columbus, Ohio, USA (2012)
13. Van Keymeulen, J., De Tier, V.: The woordenbank van de Nederlandse dialecten (Wordbase of Dutch Dialects). In: Electronic lexicography in the 21st century: thinking outside the paper.Proceedings of the eLex 2013 conference. pp. 261–279. Tallinn, Estonia (2013)
14. Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., Stanojević, M.: The Serbian Language in the Digital Age. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors), Springer (2012)
15. Wandl-Vogt, E., Declerck, T.: Mapping a traditional dialectal dictionary with linked open data. In: Proc. of the 3rd eLex Conference. Electronic lexicography in the 21st century: thinking outside the paper. pp. 460–471. Tallinn, Estonia (2013)
16. Zlatanović, M.: Rečnik govora juga Srbije: (provincijalizmi, dijalektizmi, varvarizmi i dr.). Učiteljski fakultet, Vranje (1998)
17. Zlatanović, M.: Rečnik govora juga Srbije: (provincijalizmi, dijalektizmi, varvarizmi i dr.). Aurora, Vranje (2011)