The Sixth International Conference on e-Learning(eLearning-2015), 24-25 September 2015, Belgrade, Serbia

DEVELOPMENT OF OPEN EDUCATIONAL RESOURCE (OER) FOR NATURAL LANGUAGE PROCESSING

CVETANA KRSTEV
University of Belgrade, Faculty of Philology, cvetana@poincare.matf.bg.ac.rs

BILJANA LAZIĆ
University of Belgrade, Faculty of Mining and Geology, biljana.lazic@rgf.bg.ac.rs

RANKA STANKOVIĆ
University of Belgrade, Faculty of Mining and Geology, ranka.stankovic@rgf.bg.ac.rs

GIOVANNI SCHIUMA
University of Basilicata, schiuma@arts4business.org

MILADIN KOTORČEVIĆ
University of Belgrade, Faculty of Mining and Geology, miladin.kotorcevic@rgf.bg.ac.rs

Abstract: *In this paper we present the development of an online course at the edX BAEKTEL platform named "Lexical Recognition in the Natural Language Processing (NLP)". It is based on the course of the same name for PhD studies at the University of Belgrade, Faculty of Philology. There are not many courses in Computational Linguistics (CL) on OER platforms, and there is none in Serbian either for CL or NLP. We have developed this course in order to improve this situation as it can prove useful both for linguists working in corpus linguistics and computer scientists developing NLP applications. The participant will become familiar with the use of Unitex, the corpus processing system for which many valuable resources for Serbian were already developed. This course covers a broad range of topics such as pattern recognition using regular expressions, electronic dictionaries, Finite-state automata and transducers, etc. Within the course different didactic forms were used including text, video tutorials and some useful practical exercises that should facilitate the understanding of the course subject and enable the participants to easily acquire the necessary knowledge to use the existing resources for NLP for Serbian and to develop new ones.*

Keywords: *E-Learning, Open Educational Resources, Computational Linguistics, Lexical Resources, edX*

1. INTRODUCTION

Open educational resources (OER) publicly available on the web are growing quickly and are becoming very popular both for educators and students. The growing number of OERs, besides giving a wider choice of topics, helps develop new contents adapted to local conditions in terms of cultural and educational needs. [1]

Within BAEKTEL (Blending academic and entrepreneurial knowledge in technology enhanced learning) project an open education platform is being developed, for collecting and sharing resources among academic and entrepreneurial institutions in West Balkan countries. Courses' topics cover mainly the domains ICT, geoinformatics, mining and environmental protection, geology and natural language processing, the last being in the focus of this paper.

Why Study Natural Language Processing (NLP) and Computational Linguistics (CL)? Natural language processing is the technology for dealing with human language, as it appears in everyday spoken and written communication. NLP applications have become part of our everyday experience, from spelling and grammar correction in word processors to machine translation on the web, from email spam detection to automatic question answering, from detecting people's opinions about products or services to extracting appointments from your email. [2] A short overview of the field of CL/NLP is given in section 2 of this paper.

In order to complement a lack of OER in Serbia we have decided to develop the OER course "Lexical Recognition in the Natural Language Processing (NLP)". This course is offered at the University of Belgrade, Faculty of Philology at the level of doctoral studies[1]. In Section 3 we present the development within BAEKTEL project of its OER version within the edX BAEKTEL platform.[2]

The main features of Unitex, an open access and open source corpus processing system, are presented in Section 4. Section 5 presents course content with didactic criteria and specific formats used in the OER course. Concluding remarks and plans for further research are given in section 6.

[1] http://poincare.matf.bg.ac.rs/~cvetana/Nastava/1415/nastava1415-new.html
[2] http://edx.baektel.eu/

2. COMPUTATIONAL LINGUISTICS AND NATURAL LANGUAGE PROCESSING

Computational linguistics (CL) is a theoretical discipline between linguistics and computer science concerned with understanding and modelling the written and spoken language from a computational aspect.[3]Natural Language Processing (NLP) develops methods, tools and techniques used often on the basis of theoretical results of CL, computer science, mathematics and others in order to help processing, understanding and generating human language utterances, as well as enabling various forms of human-machine interaction. It becomes very important in view of the rising amount of texts and data on the web.

The term NLP is also used to describe the function of software or hardware components in a computer system which analyze or synthesize spoken or written language. The 'natural' epithet is meant to distinguish human speech and writing from formal languages, such as computer languages e.g. Java, C#, C++, etc. [2]

Interest in NLP began in 1950 when Alan Turing published his paper entitled "Computing Machinery and intelligence," [4] but in the past ten years saw a rapid and substantial growth in the commercialization of NLP. We believe that NLP is currently a critical technology for business, because large amount of resources is produced each day, almost each moment. The vast majority of information is still expressed in natural language, so language processing has an important role in production of information.[5]

Computational Linguistics started to attract attention of researchers in Serbia more than 35 years ago when Professor Duško Vitas initiated the interest in CL as a research field in Serbia, broadened its influence, and established the group of interested researchers. The first courses at University of Belgrade were introduced in 1994: in Mathematical and CL for students of General Linguistics and students of Serbian language at the Faculty of Philology, as well as NLP graduate courses for students of Mathematics at the Faculty of Mathematics. Over the last 35 years, many resources and tools for processing Serbian have been developed within the Human Language Technologies (HLT) group that gathered researchers of different background from several faculties of the University of Belgrade. Also, a number of PhD theses and master theses have been completed in this field and several are in their final phase.

The participation of the HLT group in several international projects asserted the place of Serbian in the European family of languages when resources and tools for its processing are concerned, and opened new perspectives for their growth. [6] In 2014, HTL group founded The Society for Language Resources and Technologies, dubbed JERTEH in order to promote all branches of linguistic technology at scientific, professional and practical level.[3] The latest research results in the field of NLP for Serbian are presented in [7].[4]

Given its interdisciplinarity, NLP is not envisaged as a study programme in the Serbian higher education system. Some courses from the field are presented to students at the Universities of Belgrade and Novi Sad. They are a part of computer science, electronics, library science, linguistics and psychology studies. Faculties of Philosophy in Belgrade and Novi Sad offer courses where students can get acquainted with methods of statistical text processing (course on psycholinguistics).

At the Faculty of Mathematics, University of Belgrade there are courses providing basic mathematical knowledge necessary in the field of natural language processing (especially statistics, algebra, and logic) as well as courses on lexical analysis and text mining. The most comprehensive education is offered to the students at the Department of Library and Information Sciences at the Faculty of Philology, University of Belgrade. [8][9]

There are not many courses in CL and NLP on OER platforms, and there are none in Serbian. Coursera platform provides three courses on Natural Language Processing. These courses are offered by three universities Stanford[5], University of Michigan[6] and Columbia University[7].Topics included in these courses are syntax and parsing, language modelling and word sense disambiguation, part of speech tagging and information extraction, question answering, text summarization, collocations and information retrieval, sentiment analysis and semantics, discourse, machine translation, regular expressions, language models, text classification, and name entity recognition. All of them combine textual and video lectures with quizzes and assignments for self-evaluation.

There are also courses in Italian related to NLP[8], such as the course "Linguaggio Naturale" by Francesco Cutugno from Universita di Napoli Federico II". It aims to provide an overview of the technologies of automatic processing of written and spoken language. Another valuable educational resource entitled "Elaborazione del Linguaggio Naturale" is offered freely as wiki set of web pages[9] published by Giuseppe Attardi from Dipartimento di Informatica Università di Pisa.

3. BAEKTEL

The BAEKTEL project (http:\\baektel.eu) has the objective to foster active learning and better motivation through implementation of new technologies in the teaching process. Technology enhanced learning (TEL)

[3] http://jerteh.rs/index.php/osnivaci/?lang=en

[4] http://jerteh.rs/index.php/zbornici/?lang=en
[5] https://www.coursera.org/course/nlp
[6] https://www.coursera.org/course/nlpintro
[7] https://www.coursera.org/course/nlangp
[8] http://www.federica.unina.it/corsi/elaborazione-del-linguaggio-naturale
[9] http://didawiki.di.unipi.it/doku.php/magistraleinformatica/eln/start#corsi_affini

represents all forms of teaching and learning where ICT systems are used as media for implementing the educational process, not only in educational institutions, but also through life-long learning programs, using virtual classrooms and digital collaboration. The project is aimed at developing a network, and its technology infrastructure, for collecting and sharing open access knowledge assets among various types of institutions, both academic and entrepreneurial, in different countries.

The main task of the project is to produce course materials in various languages, but mainly in Serbian, both in video and audio format, but also in written form as parallel (multilingual) corpora of lessons and texts, supported by electronic terminological resources[10], services, and functionalities for searching and browsing of terminological resources and using them for text annotation.

The project consortium[10] consists of three of the largest state universities in Serbia (Belgrade, Niš, Kragujevac), two state universities from Bosnia and Herzegovina (Tuzla, Banja Luka), one private university from Montenegro (Mediterranean), three universities from EU (Basilicata from Italy, Iasi from Romania, Ljubljana from Slovenia), and two large enterprises, one from Serbia (NIS - Gazprom Njeft) and one from Bosnia and Herzegovina (Arcelor Mittal Prijedor).

BAEKTEL platform is developed in order to enable higher education (HE) institutions to publish their course materials (video lectures, course planning materials and evaluation tools as well as thematic content).The ICT solution includes several components:

- OER repository on local edX platform (http://edx.baektel.eu).
- BAEKTEL Metadata Portal (BMP) with metadata for all published OER within BAEKTEL network (http://meta.baektel.eu).
- Terminological web application for management, browse and search of terminological resources (work in progress).
- Web services for linguistic support (query expansion, information retrieval, OER indexing, etc.) (http://hlt.rgf.bg.ac.rs/)
- Annotation of selected resources

4. UNITEX - OPEN ACCESS, OPEN SOURCE, OPEN EDUCATIONAL RESOURCES

The course participant will become familiar with the use of Unitex, the corpus processing system for which many valuable resources for Serbian have already been developed. Unitex[11] is an open source system consisting of a collection of programs developed for text analysis by using linguistic resourcesprogramme.[11]

Unitex is based on finite-state technology. It enables application of morphological electronic dictionaries and grammars to texts for a number of different languages: French, English, Greek, Portuguese, Russian, Thai, Korean, Italian, Spanish, Serbian, Norwegian, Arabic, German, Polish and many more. For text processing simple string operations can be used, but underlying e-dictionaries enable much more sophisticated approaches. Besides simple regular expressions, graphs that represent the visualization of finite state automata (FSA) can be used for complex queries. Moreover, finite state transducer FSTs (FSAs with output) can be used for text transformation.

The concept of e-dictionaries and software that enables their effective use was developed at LADL (Laboratoire d' Automatique Documentaire et Linguistique), under the direction of its director, Maurice Gross. [12] With Unitex, user can develop electronic resources such as electronic dictionaries and grammars and apply them. Text analyses can be performed at the levels of strings, morphology, and syntax. Some of the functions are:

- developing and applying electronic dictionaries of simple words and multi-word units;
- pattern matching with queries in form of regular expressions and graphs;
- text transformations;
- processing of monolingual and bilingual texts (bi-texts in which basic segments are aligned).

Unitex is freely distributed under the terms of the Lesser General Public License (LGPL). This means that everyone can redistribute Unitex freely within the terms of the LGPL license, access the source code of all Unitex modules and reuse it.

5. LEXICAL RECOGNITION IN NLP

Course content

The main topic of the presented course[12] is natural language processing based on lexical recognition. The course focuses on processing of Serbian; however, the approach presented can be applied to other languages for which similar resources were developed.

The course consisting of 11 lessons comprises textual and multimedia educational resources accompanied with quizzes and assignments for self evaluation. The material is organized as follows:

1. The brief overview of approaches and methods used in CL and NLP. Special attention is given to differences between string oriented, statistical oriented and lexical oriented processing of texts in human languages. As an illustration, a string oriented web interface to the Corpus of Contemporary Serbian[13] is presented. [13][14]
2. Unitex corpus processing system is presented from the practical point of view: how to install it and start working with it, main steps of text

[10] http://baektel.eu/?menu=partners
[11] http://www-igm.univ-mlv.fr/~unitex/
[12] http://edx.baektel.eu/courses/UB_FIL/UB_FIL1/2015/about
[13] http://www.korpus.matf.bg.ac.rs/korpus/login.php

processing, formats used for input and output texts.

3. The concept of e-dictionaries is introduced. The specificities of e-dictionaries developed to be used by applications and not humans are stressed and contrasted to those of "traditional dictionaries" (being either in paper or digital form). The content of e-dictionaries for Serbian is given in more details.
4. The simple methods of pattern matching are presented. They are based on queries in the form of regular expressions whose basic elements are either word forms (strings) or lexical masks that refer to the content of e-dictionaries.
5. The advanced methods of text searching are introduced: morphological filters – regular expressions that enable string search at the level of characters – and graphs as a tool for visualizing regular expressions that facilitates more complex queries.
6. Advanced topics in the use of graphs are given, e.g. subgraphs and input variables. The later topics enable development of transducers for input text transformation. Methods for organizing large collection of graphs are presented. Some representative examples of graph use are shown.
7. Special types of graphs and their use are presented: preprocessing graphs, graphs for the inflection of e-dictionary lemmas and graphs for enhancement of e-dictionaries (for word forms regularly derived from lemmas already in e-dictionaries).
8. The use of contexts in graphs that shift grammars modelled by regular graphs from context-free grammars towards context-sensitive grammas. These enable construction of grammars for shallow parsing. The illustration of context use is presented by recognition of one named entity class.
9. The problems and solutions of multi-word unit (MWU) recognition are presented with emphasis on e-dictionaries of nominal MWUs, particularly their inflection that has to consider complex rules for MWU inflection in Serbian.
10. The use of powerful morphological mode is presented that enables the use of lexical resources at sub-word level, as well as the use of information from e-dictionaries for output transformations by transducers. More types of variables are introduced as well as operations on them.
11. Organization of graphs in cascades enables complex text transformation. Each graph in a cascade is a transducer that transforms a text. A graph that follows works on this transformed input. A full Named Entity Recognition System for Serbian is presented as an illustration.

Didactic criteria

Course combines different types of learning materials having in mind three basic didactic criteria [15]:

- adjusting the content to the target group,
- systematization and gradualism in the teaching process,
- connecting the theory and the practice.

To satisfy the first criterion of adjusting the content to the target group, we provided a description of the course, a list of required programs that will be used and the literature related to the topics included in the course. Whenever possible, hyperlinks to the materials are provided. If the material does not exist in e-format, a link to the list of libraries where the item can be found is provided. We also described the necessary prerequisites and competences that the students must possess.

Course lessons are structured using edX studio's structure which consists of sections, units and lessons. Lessons are designed combining different textual and audio-visual components. The last part of section is usually used for short tests that should improve student interaction activity. Course topics are presented stepwise to enable users gradually advance until the end of the course. These principles are used to satisfy the criteria of systematization and gradualism in the teaching process.

The third principle of connecting the theory and practice is fulfilled by using the video tutorials. These tutorials are designed to present practical use of processes described in theoretical parts of lessons.

Video tutorials are made by recording content of the screen and instructors voice. We concluded that this is the most suitable solution for recording videos. It is the most practical way to demonstrate the application of Unitex.

Graphical materials are used in the form of tables, pictures, print-screens... Wherever it was possible, we tried to provide interaction between the text and the picture. It was done by labelling parts of the text that is represented on the image. Those labels are matched on the image.

Some complex graphs are presented with the zooming option. Sections within the graph can be zoomed by clicking on the main view. Image 1 depicts an example of zooming usage for showing a subgraph. When the user clicks on the graph, a bubble displays the content of the subgraph Petlja01.

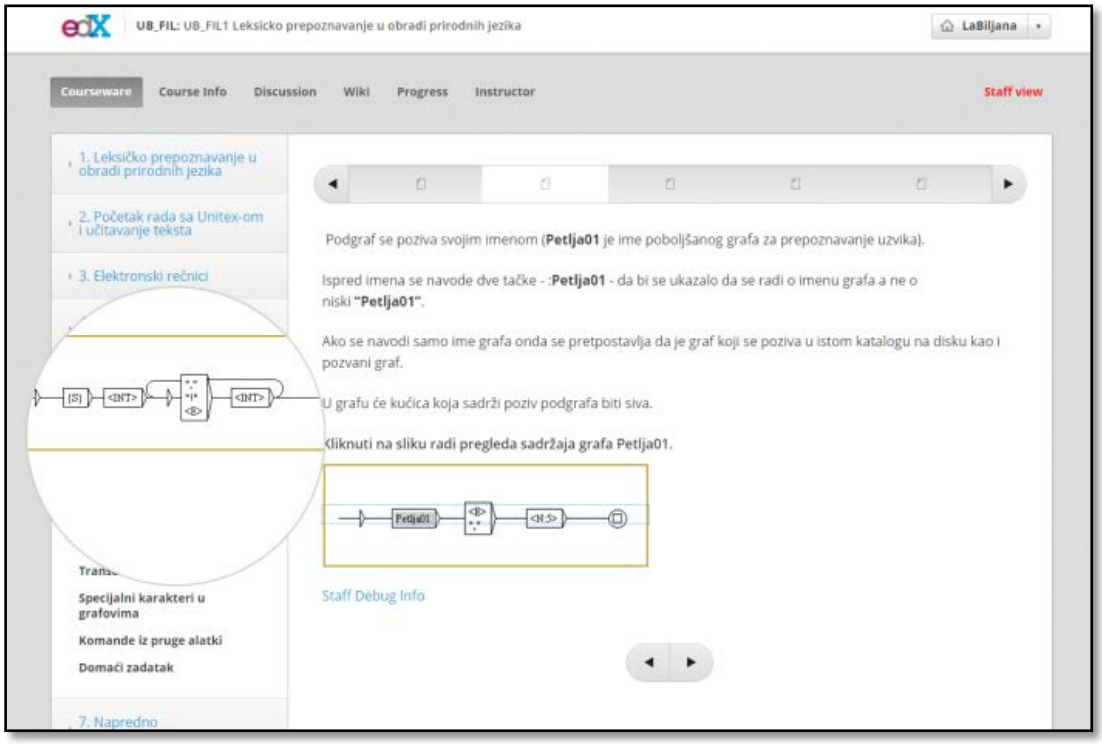

Image 1: Zooming subgraph

Tests are prepared using the problem component. To keep students' attention we used different forms for questions. We used common forms such as checkboxes, multiple choice, numerical and text input, but also advanced image mapped input and math expression input. Picture 2 represents a simple example of question where the answer requires a click on the part of a picture. After submission, the student is able to see the explanation.

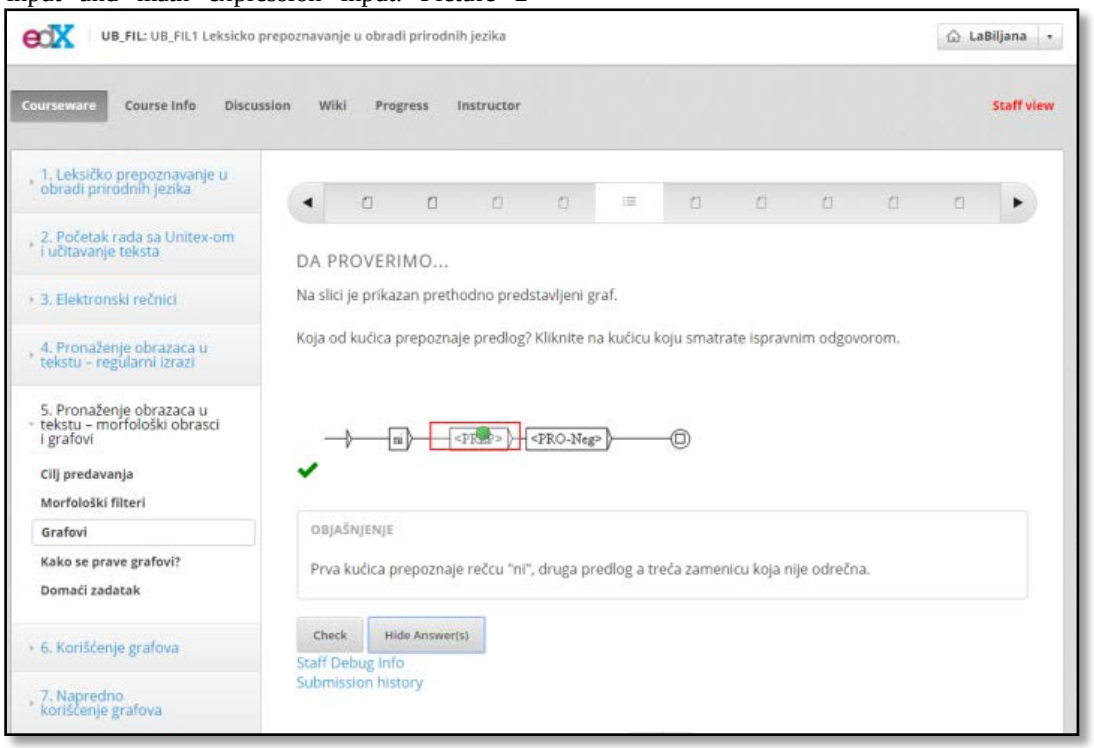

Image 2: Question - image mapped input

Specific formats

It was a big challenge to embed different formats in the course. Unitex produces and uses a variety of specific formats during the processing. Some of them, like html and txt, are readable by browsers. They are included in the course by using simple hyperlinks and it is possible to load such files while reading the text. The accessibility of these formats is important because they usually represent the input or the output for Unitex and it is important for

the user to follow the changes that are occurring with the implementation of the graphs.

Graphs and transducers are represented in grf format, a Unitex specific readable format for graph representation, and fst2 non-readable format for compiled graphs used for searching. Their visualization is represented with the use of pictures during the course. However, grf and fst2 files, as well as others, such as snt (segmented text), bin, dic and inf (files used for e-dictionaries) are zipped in a file which is downloadable as a material for practice in Unitex.

6. CONCLUSION

We hope that the developed OER for lexical recognition in NLP will be used in order to reduce the lack of similar courses. We hope that participants will easily acquire the necessary knowledge to use the existing resources for NLP for Serbian and that the number of resource users will increase significantly. We also expect that new users will contribute to improvement and enlargement of existing lexical and linguistic resources and to development of new ones. Within our future work we plan to enrich video materials by adding subtitles in other languages. We also plan to add more courses in CL and NLP at the edeX BAEKTEL platform (Basics of Theory of Formal Languages, Information Retrieval, etc.).

LITERATURE

[1] Carlucci, D., et al., A platform for management of academic and entrepreneurial knowledge, in IFKAD 2015 - 10th International Forum on Knowledge Asset Dynamics: Culture, Innovation and Entrepreneurship: Connecting the Knowledge Dots, J. Spender, G. Schiuma, and V. Albino, Editors. 2015: Bary, Italy. p. 1801-1812.

[2] Jackson, P. and I. Moulinier, Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. 2002: John Benjamins Publishing Co.

[3] Schubert, L., Computational Linguistics, in The Stanford Encyclopedia of Philosophy, E.N. Zalta, Editor. 2015, Springer.

[4] Turing, A., I.—Computing machinery and intelligence. Mind, 1950. **LIX**(236): p. 433-460.

[5] White Paper Series, H. Uszkoreit and G. Rehm, Editors. 2012, Springer: Berlin Heidelberg.

[6] Vitas, D., et al., Language Technology Support for Serbian, in The Serbian Language in the Digital Age, G. Rehm and H. Uszkoreit, Editors. 2012, Springer Berlin Heidelberg. p. 58-75.

[7] Natural Language Processing for Serbian : resources and applications 35th Anniversary of Computational Liguistics in Serbia, ed. G. Pavlović Lažetić, et al. 2014, Belgrade: Faculty of Mathematics. [6], 135.

[8] Krstev, C. and A. Trtovac, Teaching Multimedia Documents to LIS Students. The Journal of Academic Librarianship, 2014. **40**(2): p. 152-162.

[9] Krstev, C., Information Science Curriculum at the Undergraduate Studies of Library and Information Science, in Skup bibliotekara balkanskih zemalja: Saradnja obrazovanje kvalitet, A. Vraneš and L. Marković, Editors. 2002, Narodna biblioteka Srbije: Belgrade. p. 117-122.

[10] Stanković, R., et al., Building terminological resources in an e-learning environment, in The third International Conference on e-Learning. Belgrade, Serbia. p. 114-119.

[11] Paumier, S., Unitex 3.1 Beta: User Manual. 2015, Paris: Université Paris-Est Marne-la-Vallée.

[12] Gross, M., The use of finite automata in the lexical representation of natural language, in Electronic Dictionaries and Automata in Computational Linguistics, M. Gross and D. Perrin, Editors. 1989, Springer Berlin Heidelberg. p. 34-50.

[13] Vitas, D. and C. Krstev, Processing of Corpora of Serbian Using Electronic Dictionaries. Prace Filologiczne, 2012. **LXIII**: p. 279-292.

[14] Utvić, M., Izgradnja referentnog korpusa savremenog srpskog jezika. 2014, Univerzitet u Beogradu, Filološki fakultet: Beograd.

[15] Radojičić, M., et al., Creating an environment for free education and technology enhanced learning, in The Fifth International Conference on e-Learning (eLearning-2014). 2014: Belgrade, Serbia. p. 44-47.