7th International Scientific Conference
Technics and Informatics in Education
Faculty of Technical Sciences, Čačak, Serbia, 25-27th May 2018

Session 2: IT Education and Practice

UDC: 51-7

A Mathematical Learning Environment Based on Serbian Language Resources

Marija Radojičić[1*], Ivan Obradović[2], Ranka Stanković[2], Miloš Utvić[3], Sebastijan Kaplar[1]
[1] University of Novi Sad/Faculty of Technical Sciences, Novi Sad, Serbia
[2] University of Belgrade/Faculty of Mining and Geology, Belgrade, Serbia
[3] University of Belgrade/Faculty of Philology, Belgrade, Serbia
[*] marija.radojicic@uns.ac.rs

Abstract: *In recent years, in line with ever growing usage of Information technology, the learning environments are changing. The amount of available learning materials in various forms has increased. These new environments demand comprehensive learning systems, which enable management of the learning corpus with special attention paid to relevant lexical resources. In this paper we present the concept of a Mathematical Learning Environment in Serbian (MLES), which is based on a corpus of mathematical materials and various lexical resources, enabling semantic search of mathematical content. A specific use of the system is mathematical support in solving real life problems from engineering practice. To that end complex issues had to be resolved, such as mathematical text analysis, processing of mathematical content in different formats, search of mathematical materials, indexing of mathematical content using Serbian lexical resources, issues that are further complicated due to rich Serbian morphology. This paper outlines the structure and solutions for MLES, as well as the main features of its already developed components.*

Keywords: *mathematical content; text processing; mathematical formulae*

1. INTRODUCTION

Rapid development of information technology, resulting in a growing number and availability of learning materials, had a strong impact on changes in learning environments. New learning environments are needed, requiring appropriate technology to facilitate access to and management of learning materials in specific domains. Bearing this in mind, development of a Mathematical Learning Environment in Serbian (MLES) has been initiated. MLES is intended as a learning environment with the main goal of processing mathematical content in Serbian. The environment built around a corpus of mathematical content and provides mechanisms for processing and search of this content. It relies on existing lexical resources, morphological e-dictionaries and WordNet of Serbian, which have been developed within the University of Belgrade Human Language Technology group for several decades [1], as well as a newly developed glossary, Termi. The system is aimed at providing for enrichment of these resources with terms from the mathematical vocabulary, as well as offering support in understanding and solving some real life problems based on mathematical concepts. Source data for MLES corpus can vary considerably, as mathematical materials in Serbian are available in different formats, alphabets and dialects. Besides coping with different ways of writing mathematical formulae, one of the main challenges in obtaining a searchable format of these materials is the conversion of source files, given specific Serbian letters.

MLES provides a semantic search engine, which allows for processing of various formats of user requests, effective matching and relevance ranking, and features a user friendly interface. There are also possibilities of linking to BAEKTEL - an Open Educational Resources (OER) platform under edX [2], as well as relevant educational courses under a Moodle platform [3] [4].

There are several projects reported which are similar to MLES. Three interesting systems are presented in [5],[6]. MMT is a system that includes the theory graphs as the modular representation paradigm for mathematical knowledge. MathHub.info is an archive system for encoded knowledge and MathWebSearch is an example of a search engine enabling the search of mathematical formulae on the Web. The engine harvests the web for content representation of formulae and indexes them with substitution tree indexing. MathWebSearch can process only materials based on MathML and OpenMath. The project Digital Library of Mathematical Functions presents a comprehensive search mechanism for a specific corpus [7]. The corpus is based on a digitalized handbook, which contains primarily mathematical formulae, graphs, methods of computation,

references, and links to software. The search offers feedback with appropriate concordances. Mathematical formulae are converted to LaTex and then indexed. User requests are also converted to LaTex, and the search proceeds as with ordinary text. Another relevant project is MathGo! that provides search and presentation of mathematical encoded text [8]. The software solution is based on the concepts of math block identification and vector representation, with special attention paid to the search of mathematical topics using clusters and relevance ranking. The system provides ranked listing of results, through a user friendly interface, which allows seamless interaction with users through a simple query mechanism. EgoMath is yet another software solution, which allows semantic search of mathematical content [9]. It supports indexing and search of mathematical content on the web using a full text search engine. To that end the solution uses linearization, transformation rules, generalization rules and ordering algorithm, which simplify the complex and highly symbolic mathematical structures into linear structures with well-defined symbols.

This type of support for Serbian is still not available. Existing Serbian lexical resources and tools enable efficient text search, including semantic and morphological expansion of user queries, the latter being very important in highly inflective languages, such as Serbian. Of special importance is LeXimir, a tool developed within this group that greatly enhances the potential of manipulating each particular lexical resource as well as several resources simultaneously [10]. Although the resources and tools have already been successfully used for a number of various language processing related tasks including query expansion, they need further improvement for management, named entity recognition, terminology extraction, and document indexing of mathematical content.

In the next section we give and overview of the MLES system, followed by a section outlining the main issues to be solved in its development. Section Four describes corpus processing results in MLES in more detail, while section Five offers the main features of the newly developed terminological resource, Termi. The paper end with some concluding remarks.

2. FLOW CONTROL

The architecture of the MLES system is based on three main components. The first component is dedicated to corpus processing and alignment with existing lexical resources. (Figure 1).

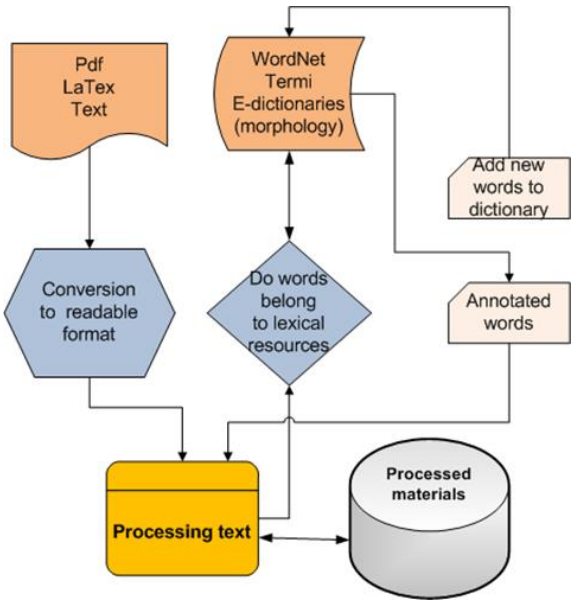

Figure 1. *Structure of corpus processing*

The obtained results are processed text, augmented dictionaries and annotated content. In this component a special challenge to corpus processing results from the use of two alphabets: Latin and Cyrillic, with different coding schemas and formats of source texts, as well as from various ways of expressing mathematical content. In order to resolve the problem of two alphabets, the entire corpus is transliterated into Latin alphabet. As for the various expression of formulas, mathematical content is converted to LaTex, which allows for expression of mathematical formulae in text only format.

The second component handles user queries, semantic search, search expansion and ranking of results. This component proceeds in several phases, such as transliteration, tokenization and lemmatization of user queries, semantic search, query expansion, expanded search and ranked retrieval results (Figure 2).

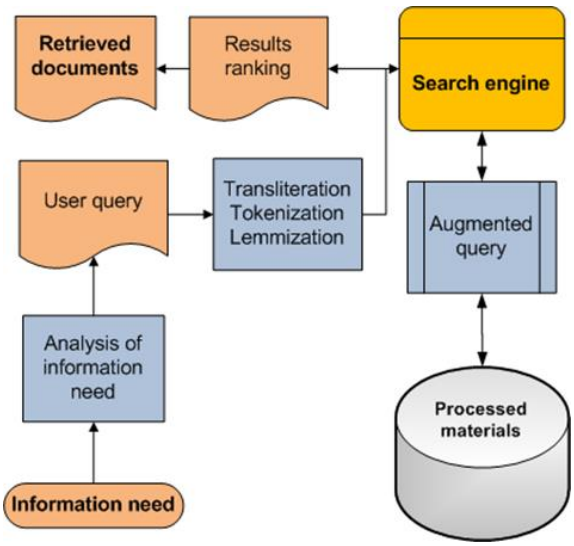

Figure 2. *Analysis of user query and semantic search*

The third component handles application to real life problems from engineering practice based on mathematical concepts (Figure 3). Results of the third component are annotated and linked texts, where every mathematical term in the text is linked to the appropriate dictionary entry or relevant corpus content related to that term.

This system component also extracts mathematical concepts from problems related to engineering practice. The process is based on clustering, categorizing and defining mathematical concepts from the base of relevant engineering problems.

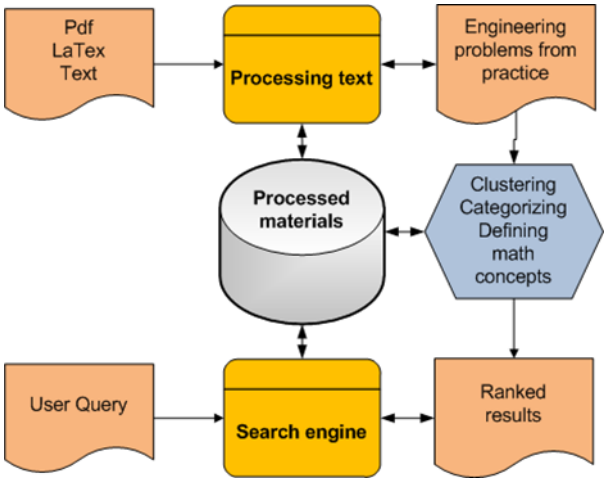

Figure 3. MLES application

3. GOALS AND CHALLENGES

Searching and processing mathematical materials is a complex problem. Standard text processors cannot recognize mathematical texts in a proper way. There is thus a need for developing new and adapting existing processors for that purpose. Processing of mathematical content requires the translation of source content into some searchable format such as MathMl or LaTeX, as a precondition for search with a search engine. MathMl is the mathematical markup language, which has the aim to integrate mathematical formulae into web pages and documents, while LaTeX notation is more in use among mathematicians in offline conditions. The idea of MLES is to convert all source materials from corpus to LaTeX format, where mathematical formulae will be presented as strings, which will facilitate processing and search of mathematical content.

One of the most important parts in processing mathematical content is semantic search of mathematical formulae. According to [5], [6] there are several challenges in searching and processing mathematical formulae, the main problem being different notation depending on the context. For instance there can exist different expressions for the same mathematical content, with the same meaning such as:

$$\frac{1}{x} = {}^{1}\!/_{x} = 1/x = x^{-1}$$

On the other hand, an expression can represent different content depending on the context, such as the number ϖ (Pi), which can present the transcendent number Pi= 3,14159 26535 89793… or radian measure of angle. Such challenges are addressed by augmented annotation and search. During the processing of mathematical formulae, augmented annotation can be realized, which can cover different expressions of the same formula.

4. CORPUS PROCESSING RESULTS

Mathematical terminology in Serbian is unsatisfactorily represented in terminological resources. Thus for example, in the Dictionary of the Serbian Academy of Sciences and Arts, out of 200,000 dictionary entries only 369 are marked as belonging to the Mathematics domain. One of the aims of MLES is to contribute to a better representation of mathematical concepts in terminological resources.

The initial MLES corpus was produced from 243 PhD theses, 15 textbooks in various areas of Mathematics as well as 212 lecture notes. Special attention was paid to the validity of content.

For the purposes of lexical processing the entire corpus is converted to textual format. As the documents were in different alphabets and encodings, as well as in different formats, a tool was created for preprocessing and normalization of texts into the Latin alphabet. In textual format the corpus contains 1,802,519 simple forms of which 118,027 are different.

Existing Serbian morphological e-dictionaries of simple forms (DELAS) and inflected forms (DELAF) contain 135,000 lemmas [11], among which only 65 are marked as belonging to the mathematical domain. There are, however, more concepts from this domain, albeit without the corresponding semantic markers. One of the tasks of MLES is to enable that these markers are added.

Processing of the MLES corpus detected 5,111 unrecognized forms with a frequency greater than 1. Among them about 1,000 grammatically correct forms were identified, and on the basis of these forms 385 basic canonical forms or lemmas were produced, using the procedure described in (Krstev 2015). Among them 191 attributes (A), 174 nouns (N), and 7 verbs (V), such as the noun "ekstremum" (extreme value), represented in the DELAS dictionary of simple forms as ekstremum, N1+DOM=Math+FLX=N1 or the verb "faktorisati" (to factorize) represented in the same dictionary as faktorisati, V21+DOM=Math+FLX=V21. Some of the inflected forms of the noun esktremum in the DELAF dictionary are:

ekstremuma,ekstremum.N:mw4q

ekstremumu,ekstremum.N:ms7q

ekstremumom,ekstremum.N:ms6q

A large number of terms in mathematics, as in other domains, are multiword expressions (MWE). Thus a procedure described in [12] has been used for semi-automatic extraction of MWEs on basis of lexical resources and local grammars developed for Serbian. Special attention is given to automatic inflectional class prediction for simple adjectives and nouns and the use of syntactic graphs for extraction of MWE candidates for termbases, their lemmatization and assignment of inflectional classes.

There were 2,900 MWE candidates extracted with a frequency over 5, covering 46,000 different forms. An example of a MWE is "diferencijalno-algebarska jednačina" (differential-algebraic equation), represented in the DELAC dictionary of compounds with the lemma:

diferencijalno-algebarska (algebarski.A2:aefs1g) jednačina (jednačina.N600:fs1q), NC_2XAXN+SIN=2XAXN(sin)

This lemma provides for recognition of the inflected forms of this compound in the corpus, such as "diferencijalno-algebarske jednačine" or "diferencijalno-algebarskih jednačina", as well as all other forms generated by the transducer (local grammar) NC_2XAXN.

Evaluation and filtering of all terms is underway, in order to generate candidate MWEs to be entered into the morphological dictionary.

Serbian WordNet (SWN) currently has 21,476 synsets [13], out of them that 232 in mathematics domain, while the Princeton WordNet has 607 in this domain, which means that at least another 375 synsets need to be added to SrpWN, such as:

ENG3013860281nimplication:4,

logicalimplication:1,

conditional relation:

ENG30-13859307-n difference:4.

The enrichment of morphological dictionaries and SWN should be complemented by content synchronization (entries and literals), as well as domain markers. In the existing dictionary only the marker +Math exists for Mathematics, but adding domain markers for specific mathematic subdomains, such as Algebra or Geometry is also planned. In addition to that, semantic markers will be developed for special functions, integrals, equations and the like.

For corpus management, we have used the IMS Open Corpus Workbench (CWB) as a collection of open-source tools [14] and an adaptation of CQPweb, a web-based graphical user interface designed specifically for CWB query processor - CQP [15]. CWB is suitable for encoding, indexing, compression and decoding large text corpora (up to 2 billion words) with multiple layers of word-level annotation. CQP is a powerful and efficient concordance system which can process query patterns specified both at the character level (specifying a form of an individual word or an annotation) and at the token level (specifying syntactic relationships between tokens). Through CQP web users can both specify query patterns and get statistical information about corpus.

Within preparation of the MLES corpus to each word within the corpus the following information is assigned, in the following order:

- Word type (noun, verb, adjective, etc.) - POS tagging
- Lemma (nominative singular for a noun, infinitive for the verb, etc.) - lemmatization
- Values of inflective categories (gender, number, case, verb form, etc.), that is, inflective base and suffixes - grammatical annotation
- Marker for the semantic value - semantic annotation.

The advantages of corpus annotation are the following:

- Corpus search becomes more efficient due to the possibility of formulating more precise queries.
- When search results are concerned, annotation.
- Compensates for the information lost during corpus preparation (removed figures, tables, footnotes, etc.), as well as information that lack due to insufficiently wide context in which the search results are presented.

Annotation also alleviates the statistical analysis of the corpus, namely automatic assignment of the distribution of annotated linguistic properties.

5. APPLICATION TERMI

The Termi application has recently been launched to serve as a support for the development of terminological dictionaries in various fields. In MLES it is used for development of mathematical vocabulary. The realization of the application was based on the ASP.NET Framework for C# programming language and MVC design pattern, as well as HTML and JavaScript, whereas SQL Server served as support for the database.

The application is located at http://termi.rgf.bg.ac.rs/ and consists of 5 specific units: browse, search, update, bibliography and profiles. Termi currently supports the processing and presentation of terms in Serbian and English, but support for other languages is also planned.

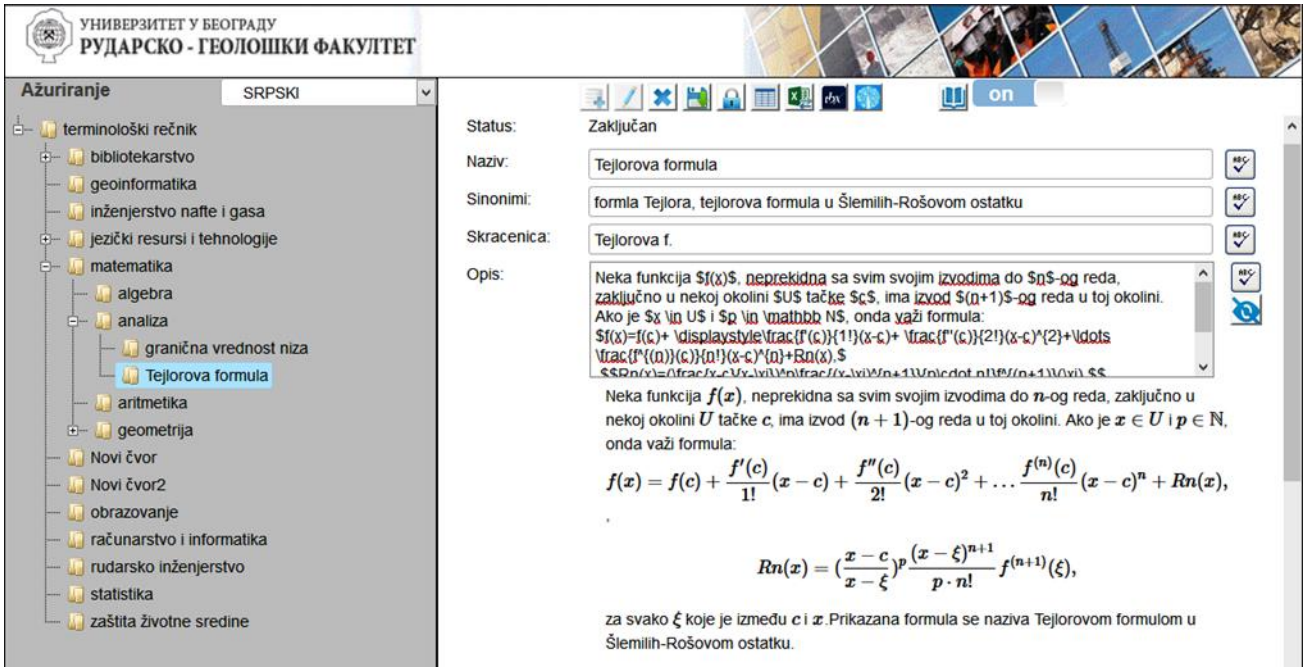

Figure 4. *Display of mathematical content through Termi editor*

On the Browse page all terms verified by editors can be viewed. The page is visible to all users regardless of whether they are logged in or not. On the left side of the page a hierarchical display of the vocabulary terms is available. Besides its name, each term has its synonyms, abbreviations, description and bibliography. In case that the description of a term contains a Latex fragment, the fragment will be interpreted, which helps in the presentation of mathematical formulae (Figure 4).

As for the Search page, it is meant for the search of terms, both in Serbian and English. This page is also intended both for users that are logged-in and those that are not.

The Update page is the most complex page in the application. This page can be accessed only by registered users, who can add, modify and delete terms both in Serbian and English. Thus, there is a possibility of updating terms either only in Serbian, or only in the English, or simultaneously in both languages. Term modification implies changes of the very properties of the term (name, abbreviation, synonyms and description) as well as modification of external connections of the term with the existing bibliography. Two more options are available on the Update page, namely spell check for the languages in which terms are entered, and the possibility that, depending on user needs, the term description is interpreted as a Latex document. On this page, akin to Browse page, on the left side of the screen a hierarchical view of the terms is available.

However, unlike the browse page, at the Update page all terms are visible, and not only terms that have been verified. In addition to update, this page offers the options of exporting data to Excel or TBX files [16]. A detail of such an export of a term is depicted in Figure 5. The Bibliography page contains a list of all bibliography units. Besides reviewing the bibliography, adding, modifying and deleting bibliography is also possible on this page.

As for the profiles, it is important to note that all logged users are divided into 4 roles: reader, editor, reviewer and administrator. The reader is a user who only has right to read, which is a role that is by default assigned to every user at registration. Editor and reviewer are users who have the task to update the contents of the dictionary, with the difference that in the hierarchy the reviewer is at a higher level. Reviewer is a user who has the exclusive right to initiate term verification (this is the advantage of reviewers in relation to the editors). As in most applications the administrator is the user who has all rights within the application. Finally, it should be noted that as a precaution logical deletion is performed not physical, while all changes are stored in a separate table. Naturally, only the administrator has the right for physical deletion.

```
<termEntry id="c110740">
    <langSet xml:lang="sr">
        <ntig>
            <termGrp>
                <term>Tejlorova formula</term>
                <termNote type="termType">entryTerm</termNote>
            </termGrp>
            <descripGrp>
                <descrip type="definition">Neka funkcija $f(x)$, neprekidna sa svim svojim izvodima do
                $n$-og reda, zaključno u nekoj okolini $U$ tačke $c$, ima izvod $(n+1)$-og reda u toj
                okolini. Ako je $x \in U$ i $p \in \mathbb N$, onda važi formula:$f(x)=f(c)+
                \displaystyle\frac{f'(c)}{1!}(x-c)+\frac{f''(c)}{2!}(x-c)^{2}+\ldots\frac{f^{(n)}(c)}{n!}(x-c)
                ^{n}+Rn(x),$$Rn(x)=(\frac{x-c}{x-\xi})^p\frac{(x-\xi)^{n+1}}{p\cdot n!}f^{(n+1)}(\xi),$$za
                svako $\xi$ koje je između $c$ i $x$.Prikazana formula se naziva Tejlorovom formulom u
                Šlemilih-Rošovom ostatku.</descrip>
            </descripGrp>
        </ntig>
    </langSet>
    <langSet xml:lang="en">
        <ntig>
            <termGrp>
                <term>Taylor's formula</term>
                <termNote type="termType">entryTerm</termNote>
            </termGrp>
            <descripGrp>
                <descrip type="definition">Let $f(x)$, and its first $n$ derivatives be continious, and let
                $f^{(n+1)}(x)$ existes on the region $U$ of point $c$. If $x \in U$ and $p \in \mathbb N$,
                than: $f(x)=f(c)+ \displaystyle\frac{f'(c)}{1!}(x-c)+
                \frac{f''(c)}{2!}(x-c)^{2}+\ldots\frac{f^{(n)}(c)}{n!}(x-c)^{n}+Rn(x),$,$$Rn(x)=(\frac{x-c}{x-
                \xi})^p\frac{(x-\xi)^{n+1}}{p\cdot n!}f^{(n+1)}(\xi),$$ For all $\xi$ between $c$ and
                $x$.Hance this formula is named Taylor's formula in Scholmic-Roche residuum.</descrip>
            </descripGrp>
        </ntig>
    </langSet>
</termEntry>
```

Figure 5. *Detail of the TBX for one term*

6. CONCLUSION

In this paper the concept of MLES, which can be a purposeful learning environment at different levels of education in Serbian, is given. The salient feature of the system is strong lexical support. Within MLES various types of lexical resources are used as well as local grammars, with the aim to provide a comprehensive and searchable learning environment. Although the general lexica in Serbian is well covered, mathematical terminology needs further improvements. MLES presents a system that supports managing and usage of mathematical content in Serbian. The ultimate goal is the integration of real life problems from engineering practice in the system. Special attention is paid on the processing of mathematical content by usage of different tools which are still under development. The concept has several advantages such as: comprehensive learning environment, development of search engines which are suitable for mathematical content, processing of mathematical content and augmentation of term base of mathematical concepts. To that end a newly developed application Termi is used, as it represents a suitable dictionary for mathematical terms.

Further plans will tackle additional integration, development and testing of lexical tools and engines in MLES. Detailed evaluation of the entire system is planned, which will provide directions for further improvement of MLES.

REFERENCES

[1] Vitas D., Popović Lj., Krstev C., Obradović I., Pavlović-Lažetić G., Stanojević M. (2012). The Serbian Language in the Digital Age, *META-NET White Paper Series*, G. Rehm, H. Uszkoreit.

[2] Edx platform, http://edx.baektel.eu/

[3] Moodle platform, http://moodle2.rgf.bg.ac.rs/

[4] Stanković, R., Obradović, I., Kitanović, O., & Kolonja, Lj, (2012a). Building Terminological Resources in an e-Learning Environment. *Proceedings of the Third International Conference on e-Learning* pp. 114-119.

[5] Kohlhase, M., Sucan, I.,(2006). A Search Engine for Mathematical Formulae. *A Lecture Notes in Computer Science*, 4120 (1), pp. 241-253.

[6] Iancu, M., Kohlhase, M., Prodescu, C. (2014). Representing, Archiving, and Searching the Space of Mathematical Knowledge. *Lecture Notes in Computer Science*, 8592 (1), pp. 26--30.

[7] Lozier, L. D., (2003). NIST Digital Library of Mathematical Functions *Annals of Mathematics and Artificial Intelligence*, 38 (1), pp. 105-119.

[8] Adeel, M., Cheung, H. S., Khiyal, S. H., (2008). Math GO! Prototype of A Content Based Mathematical Formula Search Engine. *Journal of Theoretical and Applied Information Technology*, 4 (10), pp. 1002-1012.

[9] Mišutka, J., Galamboš, L., (2008). Extending Full Text Search Engine for Mathematical. *Proceedings of Towards Digital Mathematics Library*, pp. 55-67.

[10] Stanković, R., Krstev, C., Obradović, I., Trtovac, A., & Utvić, M. (2012b). A tool for enhanced search of multilingual digital libraries of e-journals. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp. 1710-1717.

[11] Krstev, C., (2008). Processing of Serbian. Automata, Texts and Electronic Dictionaries Search Engine. *Faculty of Philology of the University of Belgrade*

[12] Krstev, C., Stanković, R, Obradović, I., Lazić, B., (2015). Terminology Acquisition and Description Using Lexical Resources and Local Grammars. *Proceedings of the 11th Conference on Terminology and Artificial Intelligence*, pp. 81-89.

Mladenović, M., Mitrović, M., Krstev, C., (2014). Developing and Maintaining a WordNet: Procedures and Tools. *Proceedings of Seventh Global WordNet Conference*, pp. 55-62.

[14] Evert, S., Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference. Birmingham:* University of Birmingham.

[15] Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*. 17 (3), pp. 380–409.

[16] Stanković, R., Obradović, I., Utvić, M., (2014). Developing Termbases for Expert Terminology under the TBX Standard. *Natural Language Processing for Serbian - Resources and Applications*, University of Belgrade, Faculty of Mathematics pp. 12-26.