An Integrated Environment for Management and Exploitation of Linguistic Resources

Ranka Stanković, Ivan Obradović

Proceedings of the International Multiconference on
Computer Science and Information Technology, pp. 287 – 294

ISBN 978-83-60810-22-4
ISSN 1896-7094

An Integrated Environment for Management and Exploitation of Linguistic Resources

Ranka Stanković
Faculty of Mining and Geology
University of Belgrade,
Djušina 7, Belgrade, Serbia
Email: ranka@rgf.bg.ac.rs

Ivan Obradović
Faculty of Mining and Geology
University of Belgrade,
Djušina 7, Belgrade, Serbia
Email: ivano@rgf.bg.ac.rs

***Abstract*—In this paper we describe two tools that form an integrated environment which can be successfully used for management and exploitation of linguistic resources. Both the tools and the resources were developed within the University of Belgrade Human Language Technology Group. The tools we describe are WS4LR, a software tool that has been developed and used for solving different tasks within the Group, and a web application named WS4QE, accompanied by several web services, that enables the solution of various tasks via the web. Besides a short description of the linguistic resources for Serbian involved, we also describe how the functions of the WS4LR tool can be used for their maintenance and development, as well as some possibilities for web query expansion offered by the WS4QE web application and the use of these expanded queries. Application of expanded queries is presented in web search, exploitation of aligned text and spatial data retrieval.**

I. Introduction

VARIOUS linguistic, that is, lexical and textual resources, are being developed within the Human Language Technology Group at the University of Belgrade (further referred to as the HLT Group), which gathers researchers in the field of Computer Linguistics from different departments. The most important and the most developed among them is the system of morphological dictionaries of Serbian (SMD). Another very important and developed resource is the Serbian wordnet (SWN), a lexical database representing the semantic network of words in Serbian. Within this group of resources, the multilingual ontological dictionary of proper names Prolex should also be mentioned.

Besides these different types of dictionaries, the Group is engaged in developing other resources, such as the e-corpus of Serbian, as well as parallel multilingual corpora, with the majority of parallel texts aligned.

The resources have been developed during several decades in the framework of various projects, and hence within different methodological frameworks. Their heterogeneity called for the development of software tools that would alleviate their maintenance and further development. Another goal was their integration, which would render the solution of various tasks related to processing of e-texts much easier. Thus a tool, with the acronym WS4LR (Workstation for Linguistic Resources), was developed. This tool, which enables synchronized use of various resources, has been successfully used for various tasks within the Group. Based on WS4LR, a web application named WS4QE (Workstation for Query Expansion) has further been developed within the Group, with the aim of supporting the development and usage of linguistic resources for Serbian via the web. In parallel with this application, corresponding web services have been developed. These services are of special interest as they can also be used independently, as separate components.

This paper outlines how linguistic resources for Serbian, developed within the HLT Group, can be further enhanced and used with the help of the WS4LR software tool and the WS4QE web application. In Section Two we will give a brief description of linguistic resources for Serbian, in Section Three the main functionalities of the WS4LR tool, and in Section Four the query expansion possibilities offered by the web application WS4QE applied on different resources: web, aligned text and spatial database.

II. Linguistic resources

In this Section we offer a brief description of three most important linguistic resources for Serbian developed within the HLT Group. Namely, the system of morphological dictionaries of Serbian, the Serbian wordnet, and parallel and aligned texts. They are the basic resources around which the WS4LR and WS4QE tools were built.

A. The system of morphological dictionaries

The format chosen for morphological dictionaries was the so-called LADL format developed within the Laboratoire d'Automatique Documentaire et Linguistique under the guidance of Maurice Gross [1]. Thus the core of the system consists of the so called DELAS dictionaries of simple words and DELAF dictionaries of their morphological forms, with more than 120.000 simple words and 1.400.000 word forms.

In addition to that, dictionaries of compounds named DELAC for the lemma dictionaries and DELACF for their

The research outlined in this paper was supported by a grant from the Serbian Ministry of Science and Technological Development

morphological forms, are also being developed. These dictionaries are composed following a similar format, but they also allow the lemma to contain non-alphabetic characters: space, dash, apostrophe etc. Without going into details, we shall only point out that the generation of morphological forms of compound words is more complex, which makes the data format in these dictionaries also somewhat more complex. As dictionaries of compounds are still under development, their dimensions are currently more modest.

The aforementioned dictionaries compose the system of morphological dictionaries of Serbian – SMD. This system of dictionaries is extremely important given the very rich morphology Serbian shares with other Slavonic languages. Besides seven cases for singular and seven cases for plural, numerous interactions take place between sounds at morpheme boundaries, producing sound changes with frequent exceptions. Thus, for example, the word "otac" (father) has the following 13 word forms: *otac, otaca, oče, očevima, očeve, očeva, očevi, ocima, oci, ocem, oce, ocu, oca.*

B. Serbian wordnet

The Serbian wordnet was developed in parallel with wordnets for Bulgarian, Greek, Romanian and Turkish in the scope of the BalkaNet project, funded by the European Commission from 2001 to 2004 [2]. The enhancement of the Czech wordnet, developed within the EuroWordNet project, was also part of the BalkaNet project. Thirteen research and scientific institutions were engaged in the project, mainly from countries where the BalkaNet languages are spoken, but also from France and the Netherlands. A national development team was formed for each language, which in the case of Serbian was the University of Belgrade HLT Group. Upon the termination of this project, the development of SWN continued, and this network to date contains about 25,000 word-sense pairs organized in a little less than 15,000 synsets.

Wordnets within BalkaNet were developed in XML format following the pattern outlined by EuroWordNet. Namely, synsets related to same (or similar) concepts in different languages were connected via the Interlingual index (ILI), integrating all BalkaNet wordnets into a multilingual linguistic data base. This feature also provides for a connection with the Princeton wordnet, the first English wordnet, which is publicly available [3].

C. Parallel and aligned texts

Although monolingual parallel texts exist, parallel texts are as a rule bilingual, composed of one original text and its translation into another language. Thus, they represent two texts having the same content, but in two different languages. The majority of parallel texts collected within the HLT Groupare are aligned, with Serbian most often being one of the languages. The procedure of transforming parallel texts into aligned texts followed two basic steps with the goal of connecting equivalent segments. In the first step parallel texts were segmented into sentences, and in the second step the sentences were aligned by means of one of the available alignment methods. The tool used for the majority of alignments was XAlign, developed within LORIA, Laboratoire Lorrain de Recherche en Informatique et ses Applications [4].

Parallel and aligned texts were grouped in aligned corpora. The HLT Group developed several aligned corpora, among them the largest one being the French-Serbian corpus which contains more than a million words [5].

III. The WS4LR Tool

With the growth of the number of resources as well as their volume and content, their maintenance and development became more and more complex. In addition to that, there was also a growing need to use several different resources for solving a particular task. Thus the development of a software tool for maintenance and integration of multiple resources was initiated. There were various problems to be solved, including different format of resources, but also different coding schemes which were used in practice and over time appeared in the resources. The result was WS4LR, an integrated and adaptable software solution, which provides for the management and manipulation of individual resources, as well as their integration [6]. A web application named WS4QE subsequently emerged, based on WS4LR, followed by corresponding web services. Its goal was to make some of the functions related to the development and use of linguistic resources available on the web as well.

As the majority of functions preformed by WS4LR and WS4QE overlap, in this section we shall describe only some of the basic functions of WS4LR related to management and development of individual resources. Integrated use of resources will be illustrated in the following section, dedicated to WS4QE, as functions of this web application are essentially also functions of WS4LR.

WS4LR provides for the management and development of the system of morphological dictionaries as well as of wordnets, where both manipulation of individual wordnets and synchronized use of wordnets for different languages are supported. In addition to that, it offers possibilities for processing and visualization of aligned texts, conversions from one coding scheme to another, and migration from one resource format to another.

Although WS4LR has mainly been used for resources in Serbian, it is not a language dependent tool. All of its functions are usable for resources in other languages provided that these resources follow the appropriate formats.

A. Dictionary management

The first system for text processing on basis of dictionaries in LADL format was Intex [7]. However, Intex could not handle the processing of texts in Unicode, whereas this coding scheme became more and more frequently used. Thus two new systems, Unitex [8] and Nooj [9] were developed, both allowing the processing of texts in Unicode, and subsequently started to suppress Intex. Although text processing in all three systems is based on dictionaries in LADL format, none of them offered possibilities for management of the content of the dictionaries themselves. Thus a module was developed within WS4LR for the entry, review and update of lemmas of simple words and compounds within

SMD. The module supports the specific features of all three solutions (Intex, Unitex, NooJ), as well as the distribution of lemmas in several dictionaries, such as the dictionary of toponyms or dictionary of proper names. This is important for practical reasons, because smaller dictionaries are easier to manage.

The most important feature of the dictionary management module is the possibility of very efficient search of dictionaries, namely of finding subsets of lemmas based on different criteria. Fig. 1 depicts a form for management of dictionaries of simple words. Within this form, lemmas can be changed, deleted and copied, and new information can be added. It is also easy to obtain a contextual menu for a lemma, with other possibilities offered (the smaller window in the picture), including the establishment of a relation with another important resource, the SWN.

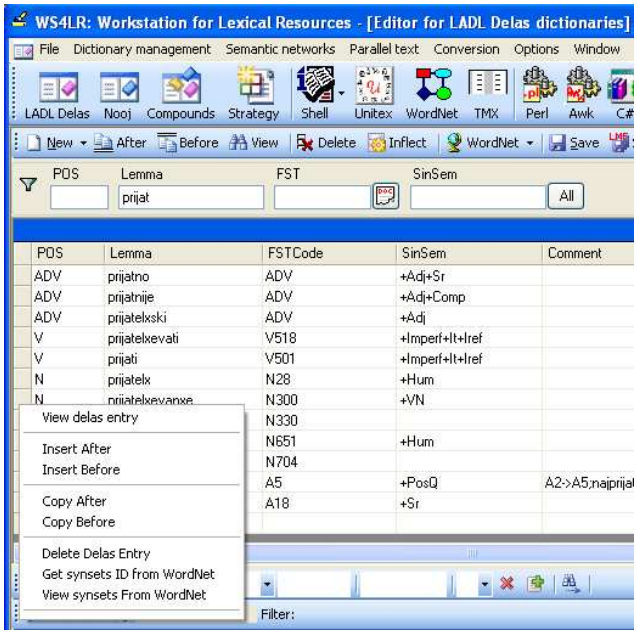

Fig. 1 Simple words dictionary management

The basic principles of search and management of data in dictionaries of compounds are the same as for the dictionaries of simple words. However, dictionaries of compounds have a somewhat more complex structure, which consequently makes their manipulation more complex. The form allowing entry of new and update of existing lemmas for compounds requires more information. Fig. 2 depicts this form for the compound "ministar za unutrasxnxe poslove" (minister for internal affairs). The upper part of the form is used for data pertaining to the compound as a whole: the code used for compound inflection, syntactic and semantic categories, comments, etc. The lower part of the form contains data related to simple word forms that are parts of the compound (inflectional class code of each simple lemma, the set of its grammatical categories, etc).

Finally, the module for management of dictionaries allows access to an editor of regular expressions, namely, transducers that describe the inflectional characteristics of a chosen lemma or class. This completes the set of tools the user needs in order to be able to manage the dictionaries.

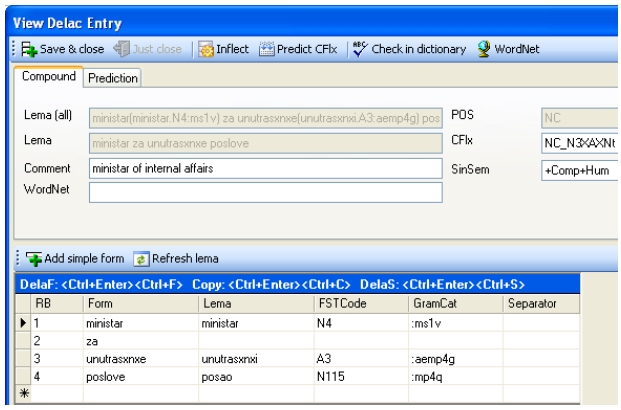

Fig. 2 Form for compound processing

B. Development of dictionaries of compounds

Development of dictionaries of compounds (DELAC) proved to be considerably more time consuming than in the case of simple words. In order to speed up this process a special component within WS4LR has been developed that assists in obtaining some of the necessary information from existing DELAS/DELAF dictionaries. Even with the help of this feature (for instance by reducing the number of errors in DELAC entries), the development of DELAC dictionaries for Serbian remained at a slow pace. Thus a procedure for automatic (or semiautomatic) construction of this type of dictionary on basis of a simple list of compounds was added.

The procedure for automatic construction of entries for a DELAC dictionary is based on a set of rules [10]. The rule design strategy results from expert knowledge on morphology and the analysis of an already existing, manually created compound dictionary. The task of the rule based procedure is to generate a complete compound lemma, based on the strategy. However, the strategy and the procedure are independent, and changes in the strategy, in general, do not affect the procedure itself. Thus experiments with various rule strategies were possible – the final strategy being the result of several iterations.

The current strategy consists of 53 rules: 19 rules for compounds with 2 components, 20 rules for compounds with 3 components, 8 rules for compounds with 4 components, and 3 rules for compounds with 5 and 6 components. Each rule defines the conditions that components of a particular compound and/or separators between them have to fulfill for a particular inflectional class to be assigned to the compound. The rules are applied in the order they are listed.

Conditions defined within each rule fall into two types: conditions of the first type specify grammatical categories of compound components and they usually apply to the components that inflect, whereas conditions belonging to the second type are additional conditions, like semantic and/or syntactic markers. Conditions are ordered by frequency in order to prioritize some conditions in case of multiple choices. For the majority of inflectional classes there is only one rule, with multiple rules defined only for a few. The or-

der of rules in the strategy reflects the probability of their application.

C. Management of wordnets

The module for management of wordnets [11] provides for manipulation of individual wordnets, as well as synchronized use of two wordnets (for example, Serbian and English), where corresponding synsets are linked via the unique identifier ILI. While working with a particular synset the user can use the hypernym/hyponym relations to ask for a tree representing the synset with parent and child nodes (Fig. 3). In that case, WS4LR also enables direct access to all synsets within the tree.

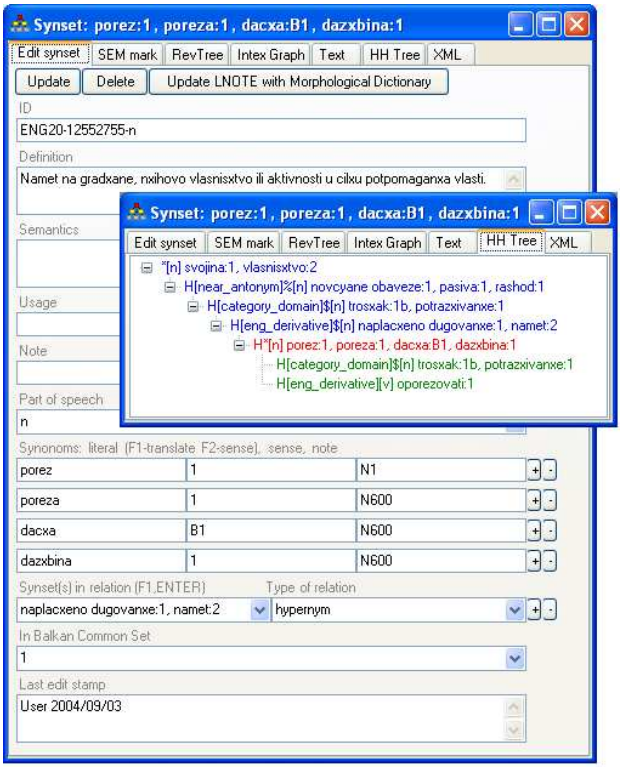

Fig. 3 Synset with the corresponding hypernym/hyponym tree

Synsets from a wordnet can be selected in much the same way as lemmas from dictionaries, using different criteria and methods, from simple entry of one or more words, to complex Xpath expressions, which can be predefined or specified by the user. The use of Xpath expression is possible because the internal representation of the wordnet is in XML. As in the case of dictionaries, this module enables update of existing synsets, but also creation of new ones. A new synset in one language (for example Serbian) may be created on basis of an existing synset in another language (for example English). In order to support this feature, the module provides access to bilingual, parallel word lists, which can help in translating synset literals from one language to another.

This module also offers different options for data consistency checking, such as detection of broken semantic relations, as the data model of the network itself does not provide referential integrity. Namely, it is possible to delete a synset from the wordnet even if there is another synset related to it. Thus a relation to a synset which does not exist any more can remain in the wordnet. The module can also check whether the same literal appears in two synsets related by a hypernym/hyponym relation, which is not allowed.

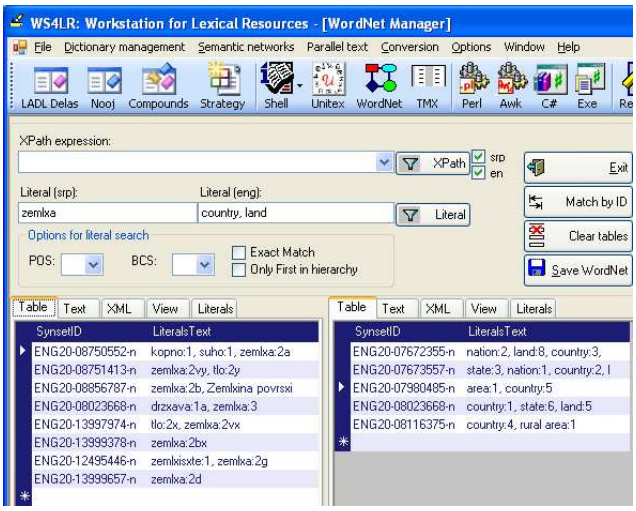

Fig. 4 The main panel for working with wordnets

Fig. 4 depicts the main panel for manipulating wordnets, showing an example where the user, during his/her search of the bilingual list positioned him/herself on the Serbian word “zemlxa”, or English “country, land”. Based on this choice, synsets that have these words in the set of their literals have been selected both in SWN (lower left side) and the English wordnet (lower right side). The user can now compare all synsets containing the literal “zemlxa” in SWN and the literals “country” or “land” in the English wordnet, and consequently perform appropriate updates and additions in SWN. This module also allows simple creation of Intex/Unitex graphs that can detect all forms of literals of a chosen synset in a given text, with the possibility of adding hypernym literals.

D. Aligned texts

WS4LR contains a module for processing of parallel texts which have previously been aligned using the text alignment tool XAlign. The module enables the transformation of texts aligned by XAlign into different formats: textual, XML, tabular or Translation Memory eXchange (TMX) format. The user can also choose the form of visualization of aligned texts. Various XSLT transformations can be applied on aligned texts in XML format, transforming them into HTML or another format, depending on the type of visualization required. The module operates on specific file structures that result from alignment with XAlign, but it can also accept as input other files that are already in TMX format.

IV. WS4QE Web Application for Query Expansion

The motivation for the development of a web application that would enable integrated use of linguistic resources in a manner similar to the one offered by WS4LR initially originated from the need to solve problems in formulating queries for web search engines. This need and the possibili-

ty to make some of the functions developed within WS4LR also available on the web led to the development of the WS4QE web application for query expansion [12]. In this Section we will offer a few examples demonstrating the possibilities our integrated environment offers in different areas of application.

A. Web search

The problems in query formulation stem from the fact that the user formulating a query is most often interested in documents related to a specific concept, in the sense concepts are defined in wordnets. On the other hand, queries for search of textual contents are usually composed of one or more keywords, connected by logical and/or operators. It is obvious that the choice of keywords, that is, their corresponding strings, is of paramount importance for the results that are going to be obtained by the query. At first glance, the main problem lies in the fact that the user might omit some of the words denoting the concept while formulating his/her query. This would reduce system recall, a parameter that represents the ratio of relevant documents obtained by the query to the total number of available relevant documents in the textual resources being searched. At first glance, this problem could easily be solved by expanding the query, namely by adding the words the user failed to use. However, the enlargement of the set of words that denote a concept in a query, although contributing to the recall in general, may also have an adverse effect. Namely, as a great number of words are polysemous, and a substantial homonymy of forms exists, the addition of new words in the query can lead to an increase of irrelevant documents in the query result, thus reducing precision, which represents the ratio of relevant document obtained to the total number of documents obtained. In view of this trade-off between recall and precision, the words or strings that are used in a query must be carefully selected in order to obtain an optimal balance between these two parameters.

The choice of strings for a query becomes additionally complicated in case of languages with rich morphological structure, such as Serbian. Namely, strings corresponding to morphological forms often need to be included in the query in addition to the basic form of the word. Some web search engines, such as Google, have attempted to solve this problem, at least partially. Thus Google recently started to expand queries in Serbian, although obviously by means of some sort of a stemmer, a program for deleting suffixes and inflectional appendices and reduction to the 'root' of the word. This solves the inflectional problem, but only partially and definitely not in a systematic way. In addition to that, the user has no control of the way Google expands the query, which can often lead to appearance of unwanted words within the query. Finally, one should bear in mind that Serbian is an example of synchronic digraphia, as two alphabets, Cyrillic and Latin, are widely used in a large variety of contexts. A query formulated in one of the two alphabets, tends to produce primarily results that contain documents in the alphabet used, which might not necessarily be the user's intention. Our tools also tackle this problem.

Lexical resources, such as electronic dictionaries and wordnets offer possibilities for a more systematic solving of the outlined problems related to query formulation. In order to obtain a balance between recall and precision, it is important to provide the user with the greatest possible flexibility in the choice of strings to be used for formulating the final query. To that end one of the basic functions of WS4QE has been developed, namely query expansion. It should be noted that a more adequate term might be query "refinement". Namely, besides query expansion, WS4QE allows the user to choose, among the strings offered, the ones to be included in the query. This means that the user can also delete strings from the expanded query if he/she estimates that they might substantially reduce precision and thus disturb the balance between recall and precision.

Various possibilities for query adjustment that WS4QE provides result from integration of several available resources. Both WS4QE and WS4LR, offer to the user the possibility to expand the query morphologically, semantically, but also to another language (English). Besides, the user can further fine-tune query formulation, since in addition to being expanded, the query can also be narrowed.

Morphological expansion is based on the use of morphological dictionaries of simple words and compounds. A possibility of morphological expansion, by a simple selection of the appropriate checkbox, is offered to the user within each form related to query formulation. If that morphological expansion is selected, WS4QE finds all inflectional forms for a given word, regardless of whether it is a simple word or a compound, using SMD, and connects these forms by logical "or" relations. In the same way the user chooses whether he/she wants the query to be in Cyrillic or Latin alphabet, or in both, by selecting the corresponding checkboxes.

WS4QE realizes semantic expansion of a query using SWN, namely, by selecting all synsets containing a given word and offering them to the user. This provides the user with an insight to all concepts the keyword pertains to, by means of a list of synonyms for each concept, as well as a definition of the concepts itself. The user then gets the possibility to delete some of these synsets if he/she so wishes, namely in the case he/she decides that they pertain to concepts which are not of interest. The query can be further semantically expanded by choosing a particular semantic relation (e.g. hypernymy/hyponymy). In that case synsets pertaining to hypernyms/hyponyms of concepts from the initial group will also appear among the synsets.

When the choice of concepts of interest is finalized, WS4QE uses them for generating a common set of words. Here again the user gets the possibility to exclude some of these words from the query. A particular word might be excluded if it is polysemous or homonymous, and its semantic relevance for the concept of interest is small, whereas its semantic relevance to another concept, which is not of interest, is considerably larger. Leaving such a word in the query would probably generate a large number of irrelevant documents. By adjusting the query with the choice of concepts and keywords, the precision of the answer obtained for the query can be considerably improved.

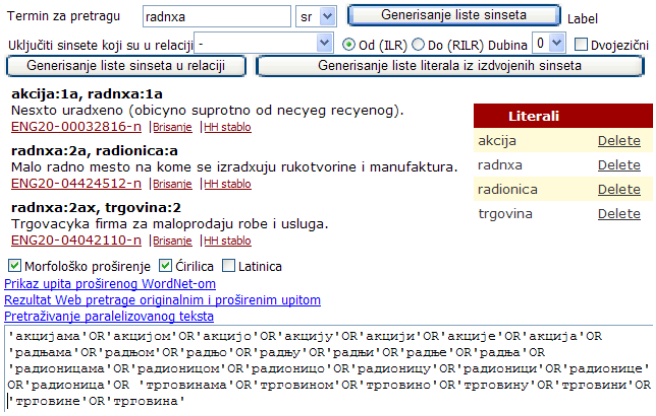

Fig. 5 Combining semantic and morphological query expansion

For example, a query submitted to Google with the compound word "političko opredeljenje" (political preference), returned a total of 24,700 documents. When the query was semantically expanded using Serbian wordnet with a synonym "ideologija" (ideology), a total of 245,000 documents were obtained. Thus the expanded query, even without further morphological expansion, obtained almost ten times more documents. However, due to the polysemy of the word "ideologija" the majority of these documents turned to be irrelevant.

The possibility of combining semantic and morphological query expansion, coupled with a choice of the alphabet, is illustrated in Fig. 5. Three synsets, with a total of four different keywords were obtained for the initial keyword "radnxa" (shop, action), each of them accompanied by a definition of the concept. The delete option is offered for each synset, but also the possibility of insight into its hypernym/hyponym tree, which can help the user decide on possible additional semantic expansion of the query. An additional list of four keywords ("literals") is generated, also offering the possibility of deleting each of them. When the user has reached his/her final decision on the keywords, he/she may then proceed to morphological expansion, and a possible change or addition of alphabet. The bottom part of the screen shows a part of the expanded query, composed of keywords or strings obtained by morphological expansion of the query previously expanded semantically, along with a change of alphabet from Latin to Cyrillic.

The third possibility is the formulation of a bilingual query, namely the addition of another language to the query. WS4QE, in the same way as WS4LR, can identify, for a given set of concepts, all corresponding concepts in another available wordnet, using the ILI. Thus, for example, for an expanded query in Serbian, one could obtain the corresponding expanded query in English, or let's say in French. This type of expansion is especially interesting if used for searching aligned texts. The formulation of a bilingual query can be combined with morphological and/or semantic expansion. Fig. 6 depicts the results of such a combination of expansions for the keyword "staklena basxta" obtained from Google. The initial query "staklena basxta" was first expanded semantically by its synonym "staklenik",

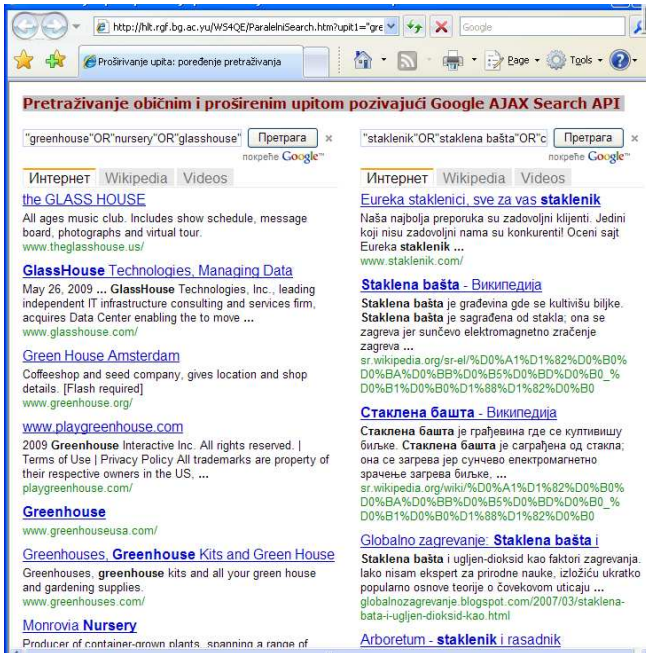

Fig. 6 Bilingual query expansion and results of Google search

and then again by adding the corresponding Cyrillic words "стаклена башта, стакленик". Then the literals "greenhouse, nursery, glasshouse" from the corresponding synsets in English wordnet were included in query.

B. Aligned text search

When a bilingual query is applied to an aligned text, WS4QE generates a filtered aligned document in TMX format. Namely, based on the expansion of the query, which can be morphological and/or semantic, segments that contain one of the word forms from the expanded query are extracted from aligned text and inserted in the filtered document. As we have already mentioned, documents in different formats, such as XML, TXT and HTML, can subsequently be generated from the TMX document filtered in this way.

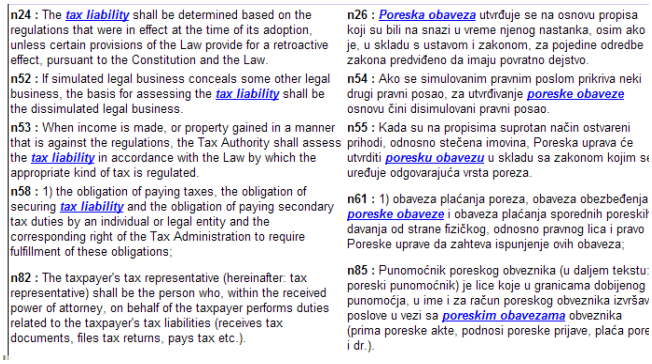

Fig. 7 Aligned segments with highlighted forms of words corresponding to a bilingual query

Fig. 7 depicts a HTML document in WS4QE with segments containing word forms from the query for the keyword "poreska obaveza" and its English counterpart "tax liability". Identified word forms are marked by underlining

and “highlighting”, namely by representing them in blue, in order to make them more easily recognizable in the text. The text in English is on the left hand side, and the corresponding text in Serbian on the right.

Given the fact that the compound “poreska obaveza” was not in the dictionary of compounds, a straightforward generation of its inflected forms was not possible. Thus the system of rules had to be invoked for determining the inflectional code of the class this compound belongs to, which subsequently entails the FST that generates its inflected forms. To that end components of this compound were individually analyzed using SMD, and the outcome was that “poreska” is an adjective belonging to the inflectional class A2 with grammatical categories*: aefs1g* (positive, same written form, feminine, singular, nominative, both animate and non-animate), and “obaveza” is a noun belonging to the inflectional class N600 with corresponding grammatical categories: *fs1q* (feminine, singular, nominative, non-animate). In this case the system found only one rule, which determined the inflectional class as NC_AXN. The next step was the invocation of the corresponding FST, which expanded the query with compound word forms: *poreske obaveze, poreskoj obavezi, poresku obavezu, poreska obavezo, poreskom obavezom,* etc. The query was further, expanded by means of synchronized wordnets to include the English word “tax liability”, but as the morphological resources for English were not available, the system was unable to perform additional morphological expansion in English, which would include “tax liabilities” as well.

Results obtained by searching aligned texts with bilingual queries can be used for different purposes, one of them being the refinement of wordnets. Namely, the analysis of aligned segments may point out to segments where translational equivalents for words in one language have not been found in the other. As the bilingual expansion is based on wordnets, the absence of a translational equivalent in Serbian indicates that the concept in question has probably not yet been included in SWN, and that its addition to SWN should thus be considered. However, in the case of the absence of a translational equivalent in English, one should bear in mind that at this moment morphological expansion is not available English. It is thus understandable why the English form ‘tax liabilities’ which is the equivalent for the Serbian form “poreskim obavezama” was not recognized in the last segment, marked as n82.

C. Spatial data search

The linguistic resources and tools presented were also implemented for improvement of spatial data search in GeolISS (Geological information system of Serbia) [13]. The system was development by the Faculty of Mining and Geology and funded by the Serbian Ministry of Environment Protection. During 2007 the Ministry supported 28 projects within various institutions aimed at gathering the necessary geological data. GeolISS is now a standard tool for handling spatial data in all projects funded by the Ministry.

Bearing in mind the vast amount of textual data in GeolISS (observations, interpretations, field work day books, etc.), morphological and semantic query expansion was introduced in the GeolISS search function. Apart from the lexical resources we have mentioned so far, semantic query expansion within GeolISS also relies on a specially developed geological dictionary, which represents a taxonomy with definitions for each entry, synonyms and bibliographical references, as well as equivalent terms and definitions in other languages (currently only English equivalents are in the database).

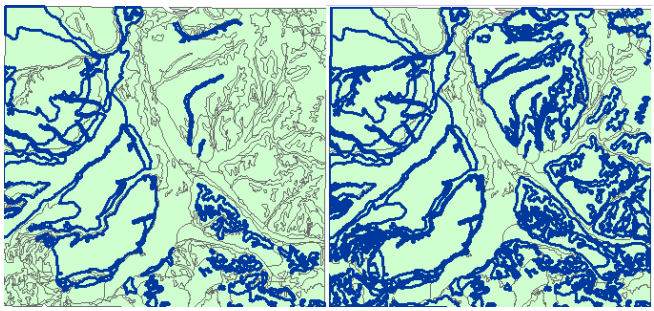

Fig. 8 Selected geologic units with the original and morphologically expanded query

For illustration purposes, a query asking for retrieval of geological units having the word “glina” (clay) in their description field was submitted twice: once without and once with morphological expansion. The word “glina” has inflected forms: *glina, glinama, glinom, glinu, glini, gline*, and the morphologically expanded query included all of them in the WHERE part of the SELECT query.

The left hand side of Fig. 8 shows the results for the unexpanded query, with 51 geologic units highlighted, whereas the results for the expanded query are on the right hand side, with 118 geologic units. The recall for the expanded query was doubled, and the precision was not reduced. Generally speaking, queries often need to be fine- tuned in order to obtain an optimal balance between recall and precision.

Semantic expansion can also be used in GeolISS to include synonyms, obtained this time from the geological dictionary, as for example: ‘nabor → bora’ (fold), ‘bezdan → ponor’ (abyss, swallow hole), ‘donji pliocen → pont’ (Lower Pliocene), ‘arteški izvor → izvor uzlaznog tipa’ (artesian spring), etc.

As an example of semantic expansion, the query asking for retrieval of geological units with the term “donji pliocen” (Lower Pliocene) in the description field was submitted twice: once without and once with semantic expansion. As “donji pliocen” has its synonym “pont” in Serbian, the expanded query introduced both forms in the WHERE part of the SELECT query. For the expanded query 33 geologic units have been retrieved (the lower part of Fig. 9), compared to 13 geologic units for the unexpanded query (the upper part of Fig. 9).

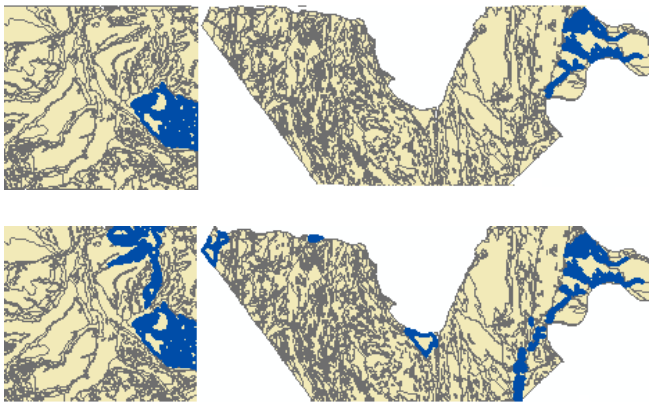

Fig. 9 Selected geologic units for the original and semantic expanded query

V. Conclusion

The results obtained so far by integration of linguistic resources justify further research on integration in several directions. In the first place, this pertains to further development of the module enabling searches with all morphological forms of compounds and free phrases. Since this search, besides the dictionary of compounds, relies also on specific rules for the generation of their inflected forms, intensive development of the inflectional module for compounds will be continued, both within WS4LR and WS4QE. Morphological query expansion for other languages (English, French…) is also planned, the main problem at this moment being the absence of adequate resources for these languages. Further enhancement of query expansion on spatial data is envisaged that will improve search results in various GIS databases.

Besides the conversions already implemented, conversion to other standard formats, such as MULTEXT-east, DCR (Data Category Registry) or MAF (Morphological Annotation Framework) are also anticipated. The implementation of derivations into WS4LR and WS4QE, which is also planned, would open a new palette of possibilities for the use of these software tools.

When the web is concerned, further development of WS4QE functions is underway, as well as the integration of developed functions and the corpus for Serbian language, which is also partly available on the web. Finally the development of a mobile application, for PDA devices and cell phones, which would enable local usage o some WS4LR functions, with the possibility of access to the WS4QE web application, is also being considered.

References

[1] B. Courtois, M. Silberztein, (eds.), *Dictionnaires électroniques du français* . Langue française 87, Paris, Larousse, 1990.

[2] D. Tufiş, (ed.), *Special Issue on BalkaNet Project, Romanian Journal on Information Science and Technology*, Bucureşti, Publishing house of the Romanian academy, 2004.

[3] P. Vossen, *EuroWordNet. A multilingual database with lexical semantic network* , Dordrecht, Kluwer, 1998.

[4] P. Bonhomme, T.M.H. Nguyen, S. O'Rourke, "XAlign: l'aligneur de Langue & Dialogue", http://www.loria.fr/equipes/led/outils/ALIGN/align.html, 2001.

[5] D. Vitas, C. Krstev, "Structural derivation and meaning extraction. A comparative study on French-Serbo-Croatian parallel texts", in *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, G. Barnbrook, P. Danielsson, M. Mahlberg (eds.), Birmingham: Univ. of Birmingham Press, pp. 166-178, 2005.

[6] C. Krstev., R. Stanković, D. Vitas, I. Obradović, "WS4LR: A Workstation for Lexical Resources", *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, Genoa, Italy, pp. 1692-1697, 2006.

[7] M. Silberztein, " INTEX: a Finite State Transducer toolbox", in *Theoretical Computer Science,* vol. 231, no.1, pp.33-46, Jan 2000.

[8] S. Paumier, "M anuel d'utilisation du logiciel Unitex", IGM, Université de Marne-la-Vallée. http://www-igm.univ-mlv.fr/~unitex/ manuelunitex.pdf , 2002.

[9] M. Silberztein, "NooJ: an Object-Oriented Approach", in *INTEX pour la linguistique et le traitement automatique des langues* . C. Muller, J. Royauté, M. Silberztein, Eds. Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté., pp. 359-369, 2004.

[10] R. Stanković, "Improvement of Queries using a Rule Based Procedure for Inflection of Compounds and Phrases", *Polibits, Special section: Natural Language Processing, Journal of Research and Development in Computer Science and Engineering,* ed. G. Sidorov (ed.), Centro Innovacion y Desarrollo Tecnologico en Computo, Instututo Politecnico Nacional, Mexico, vol. 37, pp. 14-20, 2008.

[11] I. Obradović, R. Stanković, " Wordnet Development Using a Multifunctional Tool " , *Proceedings of the International Workshop Computer Aided Language Processing (CALP) '2007* , Borovets, Bulgaria, C. Orasan, S. Kuebler (eds.), pp. 25-32, September 2007.

[12] C. Krstev, R. Stanković, D. Vitas, I. Obradović, "The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines", *in Proceedings of the Sixth Interantional Conference on Language Resources and Evaluation (LREC'08),* Marrakech, Morocco, European Language Resources Association (ELRA), May 2008.

[13] R. Stanković "Improvement of geodatabase queries within GeoIISS", *Review of the National Center for Digitization*, Faculty of Mathematics, Belgrade pp.65-74, 2008.