Softverski alati za korišćenje resursa za srpski jezik;Software tools for Serbian lexical resources

Ivan Obradović, Ranka Stanković

SOFTWARE TOOLS FOR SERBIAN LEXICAL RESOURCES

Ivan Obradović,
Faculty of Mining and Geology
University of Belgrade

Ranka Stanković,
Faculty of Mining and Geology
University of Belgrade

Abstract: In this paper we describe how lexical resources for Serbian, developed within the Human Language Technology Group, such as various types of electronic dictionaries and aligned texts, can be further refined and used for different purposes, by means of tools specially developed for these tasks. The tools we describe are WS4LR, a software tool that has already 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 lexical resources for Serbian involved, we shall 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.

1 Introduction

Development of computer linguistics, supported by rapid development of computer technology, contributed to the fact that a vast number of lexical and textual resources today are developed, stored and used in electronic, or e-format. Thus today e-texts and e-documents, sometimes bearing e-signatures are exchanged via e-mail, whereas e-dictionaries can be used for writing or analysis of e-texts.

Electronic dictionaries or e-dictionaries, taken in their broadest sense, as sets of words of a particular language systematized and organized in a specific manner, are developed in various formats. Thus, for example, several different types of e-dictionaries, along with other lexical and textual resources, are being developed within the Human Language Technology Group, which operates under the auspices of the Faculty of Mathematics of the University of Belgrade and gathers researchers in this field from several Faculties. The most important and the most developed among them is the system of morphological dictionaries of Serbian (SMD). Another highly 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 different types of e-dictionaries, the Group is engaged in developing other resources, such as the e-corpus of Serbian, as well as parallel multilingual corpora composed of parallel texts or bi-texts, usually comprising two texts of which one is original, and the other its translation. The majority of these parallel texts are aligned, which means that relations are established between corresponding elements of both texts (paragraph, sentence, word) thus yielding aligned texts.

The development of various types of resources, during a number of years and in the framework of various projects, and hence within different methodological frameworks, motivated members of the Group to initiate the development of software tools aimed at alleviating the maintenance and further development of resources, as well as their integration, which makes the solution of various tasks related to processing of texts in e-form much easier. One of the tools, with the acronym WS4LR (Workstation for Lexical

Resources), enables synchronized use of various resources and is already being successfully used for various tasks within the Group. Based on this tool, a web application is now being developed in the Group, under the "working" acronym WS4QE (Workstation for Query Expansion) with the aim of enabling the development and usage of lexical resources for Serbian via the web. In parallel with the application, corresponding web services are being developed, which are of special interest since, in principle, they can be used independently, as separate components.

In this paper we will demonstrate how lexical resources for Serbian, developed within the Human Language Technology Group, can be further developed 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 lexical resources for Serbian, in Section Three the main functionalities of the WS4LR tool, and in Section Four some possibilities offered by the web application WS4QE.

2 Lexical resources

In this Section we will give a brief description of some of the most important lexical resources for Serbian developed within the Human Language Technology Group. More precisely, about the three basic resources encompassed by WS4LR and WS4QE, namely the system of morphological dictionaries of Serbian, the Serbian wordnet and aligned texts.

2.1 Morphological dictionaries

Morphological dictionaries of simple and compound words for Serbian have been developed within the Group by C. Krstev and D. Vitas for many years (Krstev et al., 2008). The scope of the morphological dictionary of simple words is already significant, but it is nevertheless being continually developed, whereas the scope of the dictionary of compound words is at the moment more modest, which focused further dictionary development on this dictionary. The chosen format for morphological dictionaries is the so-called LADL format developed within the *Laboratoire d'Automatique Documentaire et Linguistique* under the guidance of Maurice Gross (Courtois and Silberztein, 1990). This format is widely accepted, and dictionaries in this format exist for many other languages including English, French, Greek, Portuguese, Russian, Korean, Italian, Spanish, Norwegian, Arabic, German, Polish, and Bulgarian. Entries in the morphological dictionary of simple word, named DELAS (*Dictionnaire électronique des mots simples* – electronic dictionary of simple words), are in the following form:

lema.Knnn [+SinSem]*

where *lema* in general stands for the word form of a simple word used in traditional dictionaries. *Knnn* is the so-called inflectional code which determines the inflectional class of the lemma, namely the class this lemma shares with other lemmas having the same inflectional properties. The letter *K* at the beginning of the inflectional code determines the word type, or part of speech (POS), for example, *N* for noun, *V* for verb, etc., whereas *nnn* represents the ordinal number of the inflectional class for a particular POS. The inflectional properties of the class are described by a corresponding finite automaton, or transducer[1], also labeled *Knnn*. Thus the transducer *Knnn* enables the production of all inflections, or morphological forms of the lemma belonging to the *Knnn* class. The *+SinSem* marks, which are not obligatory, but if present, may be multiple, describe the syntactic, semantic, derivational and other properties of the lemma. Thus, for example the noun *devojka (girl) appears in the* DELAS dictionary as:

devojka,N618+Hum+Ek

which means that *devojka* is a noun belonging to the inflectional class N618, denoting a human being (+Hum) and belonging to the Ekavian dialect (+Ek).

As we have already mentioned, for each inflectional class code *Knnn* a finite transducer

exists that can be used for the production of all morphological forms of the word. The generated inflectional forms of the word are stored in the morphological dictionary of simple word forms called DELAF (*Dictionnaire électronique des formes fléchies* – electronic dictionary of word forms), where the main format of data is the following:

oblik,lema[:kategorije]*

where *oblik* stands for one of the morphological forms of the simple word whose canonical form *lema* is represented in the DELAS dictionary. This is followed by :*kategorije*, namely all grammatical categories corresponding to this form, separated by a ":" sign. Thus, for example, one of the forms the transducer for the inflectional class N618 will generate for *devojka* will appear the DELAF dictionary as:

devojci,devojka.N+Hum+Ek:fs3v:fs7v

which means that two sets of grammatical categories: dative (3) and locative (7) of the singular feminine gender (f) related to a living being (v) correspond to the form *devojci* of the word *devojka*.

In addition to the dictionaries of simple words, corresponding dictionaries of compounds named DELAC for the main forms of the words and DELACF for their morphological forms, also exist. In principle, these dictionaries are following a similar format, but additionally allow *lema* and *oblik* to contain non-alphabetic characters: space, dash, apostrophe etc. Leaving the details of these dictionaries aside, we shall only stress that the generation of morphological forms of compound words is more complex, which makes the data format in these dictionaries also somewhat more complex.

All the dictionaries we have mentioned compose the system of morphological dictionaries of Serbian SMD, with more than 150,000 simple words and 1,400,000 word forms corresponding to them. The dictionary of compounds is still under development and at this moment its dimensions are more modest.

2.2 Wordnet – a semantic network of words

The semantic network of words, or simply wordnet, is based on the presumption that words, as basic elements of the language, are grouped in human mind around concepts, abstract ideas or mental symbols. Concepts encompass objects of a certain category, or a class of entities, phenomena or their mutual relationships. Different semantic relations exist between concepts. One of the basic semantic relations is the hypernym/hyponym relation which relates a more general concept to a concept representing "one of its specific categories" (e.g. animal/dog). Another frequent relation is the holonym/meronym relation which relates one concept to another concept representing one of its parts (e.g. hand/finger). There are, of course, numerous other semantic relations, and thus concepts with their semantic relations build a semantic network. In every particular language, concepts are lexicalized by one or more synonymous words, simple or compound. Thus, for example, for the concept defined as "an actor's portrait of a person in a play", the word "uloga" but also the words "lice" and "lik" are used in Serbian. Thus a semantic network of concepts in a particular language becomes a corresponding semantic network of words, which is further materialized as a lexical data base of a specific structure.

Development of wordnets started in 1985 in the research team of the *Cognitive Science Laboratory* at Princeton University, under the guidance of the renowned professor of psychology, George Miller. The first wordnet was developed for the English language under the name of Princeton Wordnet (PWN), as a linguistic database whose organization, or structure, was supposed to mimic the manner in which human brain stores and uses linguistic terms (Fellbaum, 1998). The aim of Miller's team was to create PWN as some sort of a "mental lexicon" which could be used in psycholinguistic research projects. PWN, the lexical data base that materializes the semantic network of concept for English, is based on the

representation of each concept by a set of synonymous word-sense pairs that represent the basis for the central element of this base, namely the synset. The use of the word-sense pair is based on the approach used in standard dictionaries of spoken languages, where each word may have several senses, marked in a specific manner. In the wordnet database, each synset contains, besides the word-sense pairs themselves, other data, of which the most important are the word type, or part-of speech – POS, the definition of the concept, examples of the use of words to denote the concept, and semantic relations which relate this synset with other synsets. At the end of 2007, PWN contained about 155,000 words organized into more than 117,000 synsets with approximately 207,000 word-sense pairs.

The wordnet structure developed within PWN has later been used for the development of a large number of wordnets for other languages. Some of these networks were developed within larger, multilingual linguistic data bases, in the scope of international projects for simultaneous development of wordnets for several languages. The first project that introduced multilingualism in wordnets was EuroWordNet. Besides the wordnet for English, in the scope of this project, corresponding wordnets were developed for seven other European languages: Dutch, Italian, Spanish, French, German, Czech, and Estonian (Vossen, 1998). All networks within EuroWordNet were developed following the pattern established in PWN, but EuroWordNet also introduced an important novelty. Namely, in EuroWordNet relations were established among synsets representing the same concept in different languages via an Inter-Lingual-Index or ILI. This enabled the interconnecting of wordnets for different languages and their integration into a multilingual linguistic data base. Wordnets for Bulgarian, Greek, Romanian, Turkish and Serbian were developed along the same lines between 2001 and 2004 within the scope of the BalkaNet project, funded by the European Commission (Tufiş, 2004). Within the same project the Czech wordnet, initiated within the EuroWordNet project, continued to be developed. Thirteen research and scientific institutions were engaged in the BalkaNet project, mainly from countries where the BalkaNet languages are spoken, but also from France and Netherlands. A national development team was formed for each language, and in the case of Serbian this team was the Human Language Technology Group at the University of Belgrade. Upon the termination of this project, the development of SWN continued, and this network to date contains more than 20,000 word-sense pairs organized in more than 14,000 synsets.

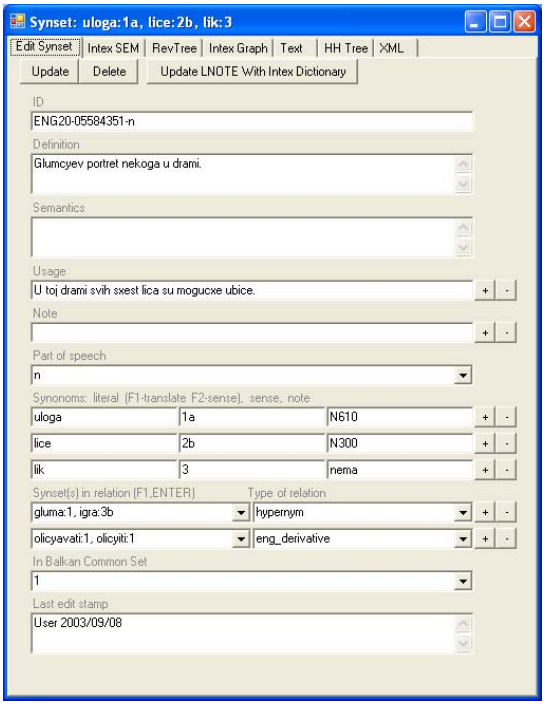

Figure 1. Example of a synset

Figure 1 depicts the SWN synset for the concept "an actor's portrait of a person in a play" as seen in the WS4LR tool. Without going into details, we shall only mention that in the development of SWN, for reasons of flexibility, the so called Aurora code has been used, where letters specific for the Serbian language are coded by two letters of the English alphabet (ć, č, š ,ž ,đ,

dž, lj and nj are coded as cx, cy, sx, zx, dx, dy, lx and nx, respectively). Among other things, this enables the use of SWN both in Cyrillic and Latin alphabet, as needed.

2.3 Aligned texts

As mentioned before, a parallel text is usually composed of two texts one of which is the original, whereas the other is generated as its translation. Thus, the majority of parallel texts are bilingual, namely, composed from two texts of the same content in two different languages. However, parallel texts can also be generated in other ways. Namely, when fiction is concerned, it often happens that different translations of the same text to a particular language exist, their appearance usually separated by a period of several years. In that case it is possible to generate a monolingual parallel text, from two texts in the same language, which in general also have the same content, because they were generated by translating the same original, but which are not identical. Finally parallel texts can consist of several texts of the same content in several languages. Such parallel texts originate from one original text and its translations in two ore more languages, and are called multilingual parallel texts.

In the majority of cases, parallel texts are being aligned, which turns a parallel texts into an aligned text. Sometimes, it is even considered that parallel texts are the same as aligned texts, but this does not always have to be the case, since non-aligned parallel texts are also sometimes being used (Ohmori and Higashida, 1999). The procedure of transforming a parallel text into an aligned text consists of two basic steps. In the first step parallel texts are split into segments, that is, basic units of text. Usually, sentences are chosen for segments, but segments can be larger, such as paragraphs, or smaller, such as words. The second step is the alignment of segmented parallel texts by means of one of the available alignment methods. The goal is to connect equivalent segments in two or more parallel texts. The method usually used for alignment at the sentence level, which is the most common one, is the method developed by (Gale and Church, 1993). Figure 2 depicts an example of an aligned text represented in the WS4LR tool. It is a legal texts in English and Serbian, aligned at the sentence level.

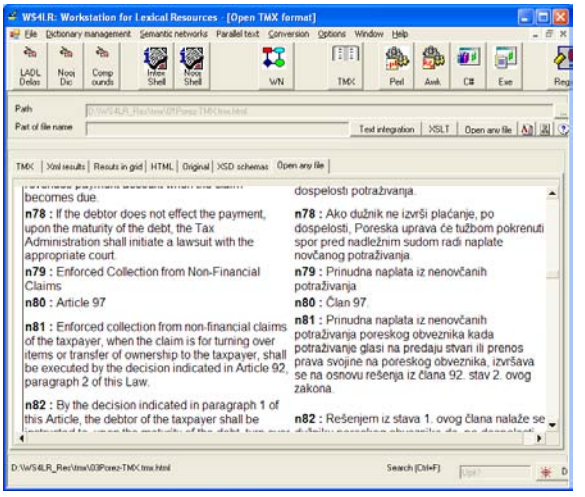

Figure 2. Example of an aligned text

Parallel corpora are very useful in the research pertaining to bilingual but also multilingual lexicography (Steinberger et al., 2006), for learning foreign languages, for translation, for linguistic research or comparative analysis of two or more languages, terminology extraction, etc. Monolingual parallel texts are especially interesting in research related to paraphrasing (Barzilay i McKeown, 2001).

The Human Language Technology Group developed several aligned corpora, among them the largest one being the French-Serbian corpus which contains more than a million words (Vitas and Krstev, 2005).

3 WS4LR – a tool for maintenance and integrated use of lexical resources

With the growth of the number of resources as well as their volume and content, a need emerged for the development of software tools which would alleviate their maintenance, use and integration and enable their further development efficiently. Several problems had to be solved, such as different format of resources, but also differ-

ent coding schemes which are used in practice and over time appeared in the resources, starting with the Aurora code, followed by the ISO 8859-2 and ISO 8859-5 codes, up to Unicode. Thus the WS4LR software tool was created, which represents and integrated and adaptable software solution, enabling the management and manipulation of individual resources, as well as their integration (Krstev et al., 2006). Based on this tool a web application named WS4QE is now being intensively developed, together with appropriate web services, in order to make some of the functions related to the development and use of lexical resources available also on the web. In this section we shall only describe some of the basic functions of WS4LR related to individual resources. Integrated use of resources will be illustrated in the section on the WS4QE web application. It should, however, be noted that all the possibilities offered by WS4QE, which are going to be described in the next section, are essentially also functions of WS4LR, since WS4QE practically represents a web upgrade of functions already developed in WS4LR.

3.1 The functional model and characteristics of WS4LR

WS4LR is a modularly organized system with the aim of providing the following functions:

– management of the SMD system of morphological dictionaries,

– development and enhancement of the Serbian wordnet SWN, in such a way as to support both the manipulation of individual wordnets, as well as synchronized use of wordnets for different languages,

– usage and presentation of aligned texts,

– conversions from one coding scheme to another, and

– conversions form one resource format to another..

We will describe here briefly the WS4LR modules, pointing out that detailed instructions for the users are available both in printed format and through on-line help which is an integral part of this software.

Although WS4LR has mainly been used for Serbian, its usage is not language dependent. The only precondition is that resources exist, that is, that they are being developed in the appropriate formats.

3.2 Management of dictionaries

Initially, the Intex system (Silberztein, 1993) was used for text processing using dictionaries in LADL format. But since Intex did not enable processing of texts in Unicode and this coding scheme became more and more frequently used, the Unitex[2] and Nooj[3] system were developed, both allowing processing in Unicode, and subsequently started to suppress Intex. Although all three systems enable text processing 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, which supports the specific features of all three solutions (Intex, Unitex, NooJ).

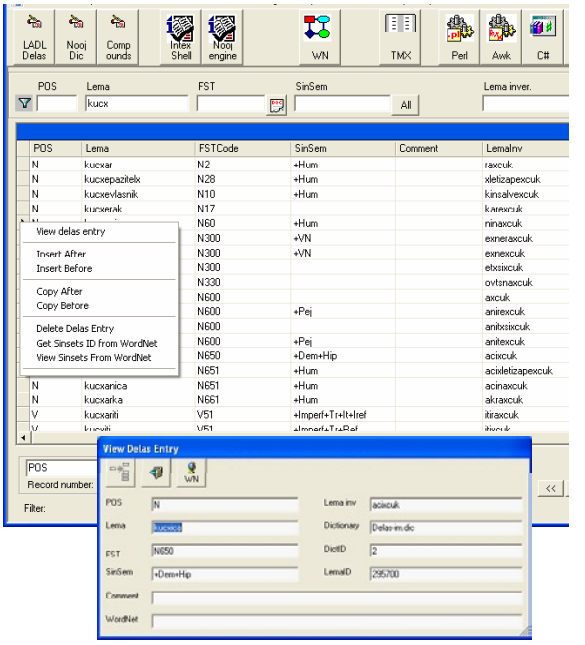

Figure 3. A panel for management of dictionaries of simple words

The system of morphological dictionaries supports the distribution of the dictionaries, namely enables 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. Besides, and which is more important, the usage of all dictionaries during text processing using the Intex/Unitex systems is not always necessary, and sometimes even advisable.

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. Lemmas can be selected according to the criterion of matching a lemma substring with a given string, then by specifying the part of speech, the inflectional class, syntactic or semantic tags, as well as by combining all these criteria through Boolean expressions. Figure 3 depicts a panel for management of dictionaries of simple words. The lemmas which have been selected and represented in table form are those starting with the substring "kucx"[4]. The lemmas can be changed, deleted or new information added to them using the depicted panel. It is also possible to add new rows into the table either by entering all elements of the lemma from scratch, or by copying one of the existing lemmas, and then performing the necessary modifications, which can often make the work easier and speedier. It is also easy to obtain a context menu for a lemma leading to other possibilities (shown with an arrow in the picture), and which enables the establishment of a connection with another important resource, the SWN.

Dictionaries of compounds have a somewhat more complex structure, which consequently makes their manipulation more complex, although the basic principles of search and management of data are the same as for the dictionaries of simple words. The form for entry of new and update of existing lemmas for compounds requires the entry of more information. Figure 4 depicts this form for the compound "gross national product". The upper part of the form is used for the entry or display of information 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 information assigned to simple forms that are components of the compound (inflectional class code of the simple lemma, the set of grammatical categories of the simple form which is a part of the compound lemma, etc).

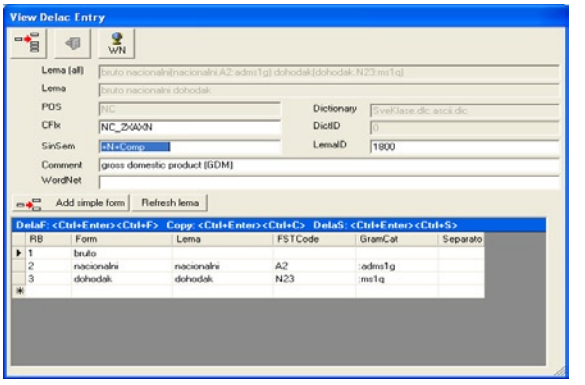

Figure 4. Form for compound processing

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

3.3 Management of wordnets

The module for management of wordnets enables the manipulation of particular wordnets, but also 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 (Figure 5). In that case, WS4LR also enables direct access to all synsets within the tree.

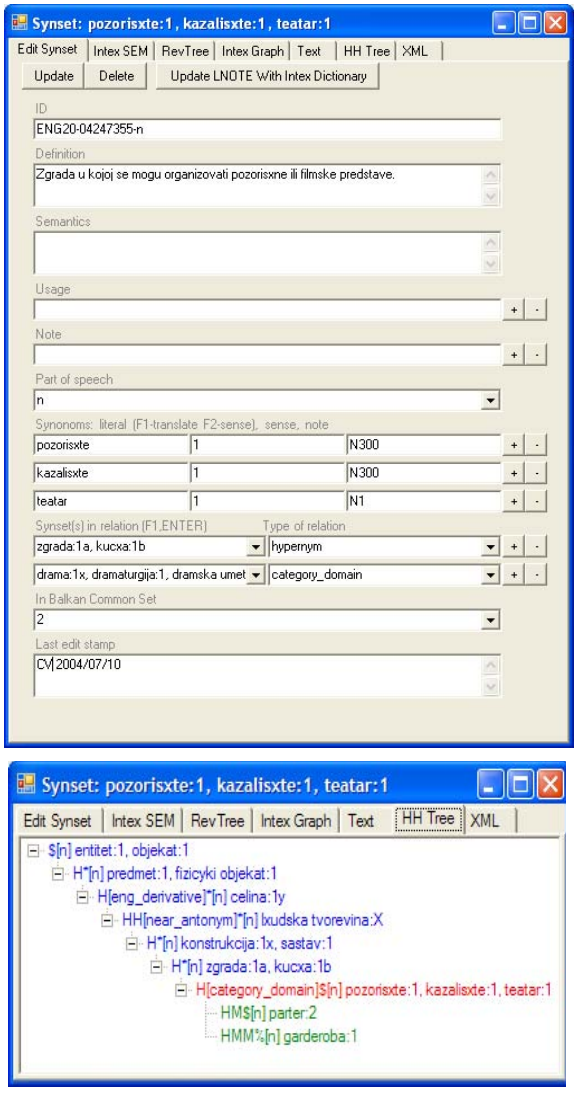

Figure 5. Synset with the corresponding hypernym/hyponym tree

Synset form a wordnet can be selected in much the same way as lemmas from dictionaries, using different criteria and methods, from simple stating of words, or more precisely strings corresponding to words, to complex Xpath expressions, which can be predefined or specified by the user. The Xpath expression can be use 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 enables the usage of bilingual, parallel word lists which can help in translation of synset literals in one language synset literals in another language.

Different options for data consistency check are incorporated in this module, such as those for detecting broken semantic relations, since the referential integrity of data is not defined in the network itself, as a data model. Namely, it is possible to delete any synset in the wordnet, regardless of whether there is another synset related to it, so a situation can occur where a relation is established with a synset which does not exist any more, or does not exist yet. Besides, the same literal should not appear in two nodes related by a hypernym/hyponym relation, which can also be verified by this module.

Figure 6 depicts the main panel for working with wordnets, using an example where the user, during his/her search of the bilingual list based on the criterion that the literal in Serbian starts with "kucx" (upper right corner), positioned him/herself on the word "kucxa", that is "house". Based on this, synsets which contain these words mong their literals have been selected both in SWN (lower left side) and the English wordnet (lower right side). The user can now compare synsets that contain the literal "kucxa" or "house" and consequently perform appropriate updates and additions in SWN.

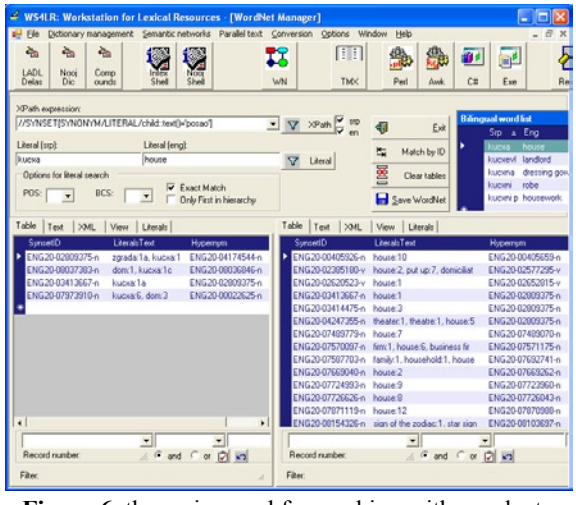

Figure 6. the main panel for working with wordnets

This module also enables easy creation of Intex/Unitex graphs which detect all forms of literals of a chosen synset in a given text, with the possibility of adding hypernym literals. The left side of Figure 7 depicts a WS4LR panel where the textual form of the graph for the synset {kuća:6, dom:3} is generated, together with its hypernyms, whereas the right side depicts the graph generated in the Intex environment. In the Intex/Unitex environment the <kucxa> represents all inflectional forms of the word *kucxa* – it is presumed that *kucxa* is a lemma in the DELAS dictionary, with all its inflectional forms generated in the DELAF dictionary.

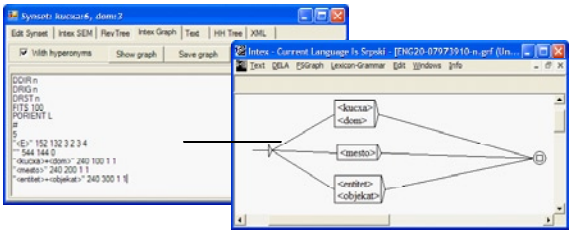

Figure 7. Generating a graph for a synset and its hypernyms

3.4 Aligned texts

WS4LR contains a module for processing of parallel texts which have previously been aligned using the text alignment tool XAlign (Bonhomme et al., 2001). The module enables the transformation of texts aligned by XAlign into different formats: textual, XML, tabular or Translation Memory eXchange (TMX) format[5]. Besides, 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. Besides the possibility of working with specific file structures which are result from the alignment with XAlign, this module also accepts other files already in TMX format as input. The panel in Figure 8 depicts an aligned text in TMX format.

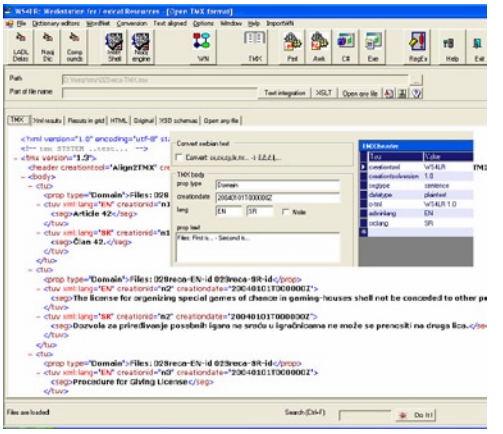

Figure 8. TMX format of aligned text

3.5 Conversions

As mentioned before, due to the fact that resources have been developed during many years, resources exist in different formats (Intex, Unitex, NooJ). Besides, since the resources have been developed in the Aurora code, they need to be transformed in other coding schemes in order to be used in "real" texts. Thus a module was developed within WS4LR enabling the user to perform conversion from one coding scheme to another, as well as from on format to another. The user can define which subset of resources should be processed, for example, a selection of dictionaries. The user can select the program code (script) he/she finds the most suitable for conversion, depending on the form of resources to be converted (text, graph, morphological dictionary, and the like), as well as on the specific demands of the conversion. The conversion is implemented mainly in the C# programming language, but external Perl or awk scripts can also be used, which enables the user to add them him/herself to the system if needed, and thus adjust additionally the conversion to his/her specific needs. It is, for example, possible to perform conversion of XML documents leaving XML tags intact, which is especially important in conversions to Cyrillic alphabet, as well as in converting graphs. When conversion of resource format is concerned, then it is most often the case of a transformation of resources such as dictionaries, graphs and regular expressions from a format

used by Intex to a format used by NooJ. Figure 9 depicts a panel for conversion of a morphological dictionary into Unicode using a C# procedure, with additional parameters of conversion specified. This module also enables conversion into LMF (Lexical Markup Framework) format (Francopoulo et al., 2006).

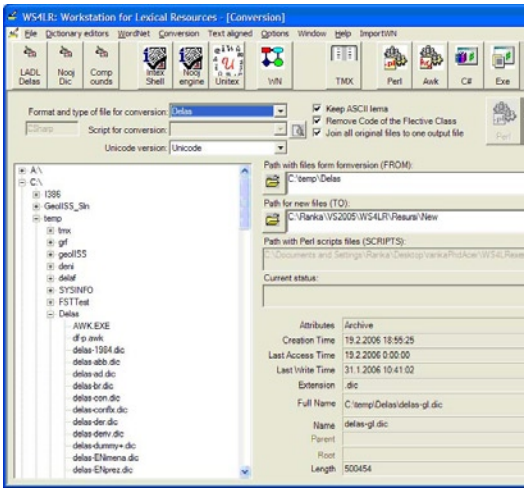

Figure 9. Panel for conversion of resources

4 Development of the WS4QE web application for lexical resources

The possibility and the need for some of the functions developed within WS4LR to become also available on the web led to the development of the WS4QE web application for lexical resources. This application is still under development[6], but some of its functions can already be used. The main menu of WS4QE is depicted in Figure 10. Due to the need to allow different modes of usage to different users, the creation of user accounts is envisaged, along with logging in prior to usage of this application. The largest set of envisaged user functions is related to query expansion, functions that can already be used and which are going to be described in more detail in this section. A group of functions related to manipulation of aligned texts is also envisaged, and since some of these functions are already available, they shall be discussed here as well. Finally, WS4QE should enable the manipulation of lexical resources (for the time being only the update and search of SWN is envisaged), and also offer information of the Human Language Technology Group and the developed software for lexical resources, namely WS4LR and WS4QE.

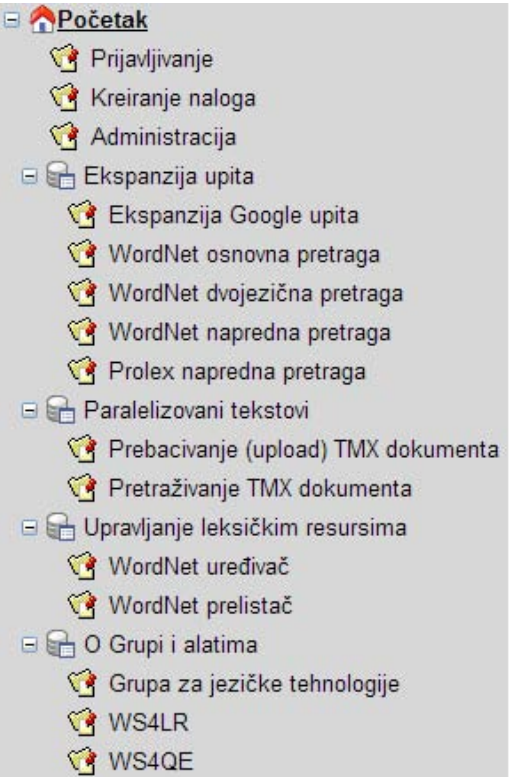

Figure 10. The main menu of the WS4QE web application

The motivation for the development of a web application which will enable integrated use of lexical resources in a manner similar to the one offered by WS4LR initially originated form the need to solve problems with formulating queries for web search engines. Thus the "working" acronym of this web application – WS4QE (Workstation for Query Expansion).

The problem of query formulation stems from the fact that the use formulating the query is most often interested in documents related to a specific concept, the way concepts are defined in wordnets. On the other hand, queries for search of textual contents are usually composed of one or more keywords, or more precisely strings corresponding to those words, 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 the 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 the query, although contributing to the recall in general, may also have an adverse effect. Namely, as a great number of words is 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 ration of relevant document obtained to the total number of documents obtained. In vies 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. Besides the basic form of the word, strings corresponding to its morphological forms often need to be included in the query. Some we search engines, such as Google, have attempted to partially solve this problem. Thus queries in Serbian recently started to be expanded, however obviously by the use of some sort of a stemmer, a program for deleting suffixes and inflectional appendices and reduction to the 'root' of the word (see Kešelj and Šipka in this number). This solves the inflectional problem only partly and definitely not in a systematical manner. As is often the case with stemmers, besides inflectional forms the stemmer often generates related words. Thus a query using the word *prevodilac* (translator) also offers answers containing the word *prevod* (translation), whereas the query with the compound *slati poruku* (send a message) produces only web pages containing the verb *slati* (send) in infinitive or the verbal noun *slanje* (sending) and omits numerous pages which contain other forms of this verb such as, for example, *šaljem poruku* (I am sending a message). In addition to that the user has no control on the way Google expands the query. Finally, when Serbian is concerned, there is also the problem of two alphabets, Cyrillic and Latin. If a query is formulated in one alphabet then the query results will contain only documents in the alphabet used, which might not necessarily be the user's intention.

Lexical resources, such as electronic dictionaries and wordnets offer possibilities for a more systematical solving of the outlined problems related to query formulation. In this task, it is important to offer the user the greatest possible flexibility in the choice of strings to be used for formulating the final query, in order to obtain a balance between recall and precision. 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 "adjustment". Namely, besides query expansion, a possibility of choosing the strings to be included in the query is offered to the user, which includes possible deletion of strings from the expanded query if the user estimates that they might substantially reduce precision and thus disturb the balance between recall and precision.

Various possibilities for query adjustment that WS4QE provides are a result, as in the case of WS4LR, of an integration of several available resources. WS4QE, in the same way as WS4LR, offers to the user the possibility to expand the query morphologically, semantically, but also to another language (English). Besides, WS4QE provides further possibilities for the user to control the query formulation, since in addition to expansion, it also offers a narrowing of the query.

Morphological expansion is based on the use of morphological dictionaries of simple words and compounds. A possibility of morphological expansion, by a simple choice of the appropriate field, is offered to the user with each form of query formulation. In that case WS4QE finds all inflectional forms for the given word, regardless of whether it is a simple word or a compound, using SMD, and connects them by logical "or" relations. In the same way, by a simple choice of appropriate fields, the user selects whether he/she wants to make a query in Cyrillic or Latin alphabet, or in both.

WS4QE obtains semantic expansion of a query by means of SWN, 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, through sets of synonyms used for these concepts, as well as a definition of the concepts themselves. 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 the choice of a particular semantic relation (e.g. hypernymy/hyponymy), in which 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 excludes some of these words from the query. The motivation for excluding one of the chosen words could lie in the fact that its semantic relevance for the concept is small, while at the same time it could generate a large number of irrelevant documents, due to the fact that being polysemous or homonymous, its semantic relevance to another concept, which is not of interest, is considerably larger. By adjusting the query with the choice of concepts and keywords, the precision of the answer obtained for the query can be considerably improved.

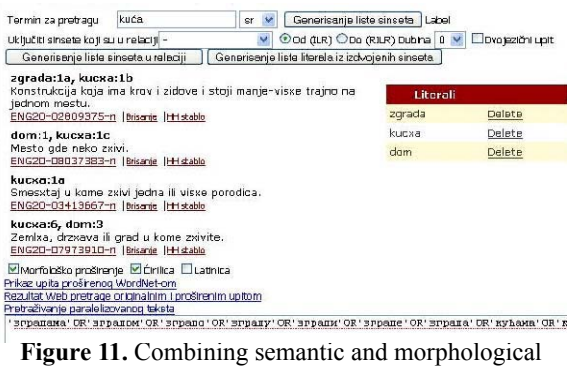

Figure 11. Combining semantic and morphological query expansion

The possibility of combining semantic and morphological query expansion, coupled with a choice of the alphabet, is illustrated in Figure 11. For synsets, with a total of three different keywords were obtained for the initial keyword "kuća" in Latin. Each synset is accompanied by a definition of the concept, as well as a possibility of its deletion, but also of an insight into its hypernym/hyponym tree, which can help the user decide on possible additional semantic expansion of the query. A list of three keywords ("literals"), with a possibility of deleting each of them, is generated from these four synsets. 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 which was 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, for a given set of concepts can identify 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 bi-

lingual query can be combined with morphological and/or semantic expansion. Figure 12 depicts the bilingual expansion for the keyword “beli luk” (garlic), with corresponding semantic and morphological expansions, as well as the results obtained by submitting such an expanded query to Google. Although the Latin alphabet was used for the initial keyword, Cyrillic was chosen for query formulation, and not Latin, and thus the results for the Serbian part of the query obtained from Google contain only pages in Cyrillic.

When such a bilingual query is applied to an aligned text, WS4QE generates a filtered aligned document in TMX format. Based on the expansion of the bilingual query, which can be morphological and/or semantic, segments are extracted from aligned text that satisfy the query, namely segments where one of the forms of the word contained in the expanded query was found. From a TMX document filtered in this way, output documents can be further generated in different formats, such as XML, TXT and HTML, as we have already mentioned.

Figure 13 depicts a HTML document in WS4QE with selected segments where one of the forms from the expanded query for the keyword ‘igra’ (game) is found in at least on of the languages. The forms identified 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.

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. Since wordnets have been used for bilingual expansion, the absence of translational equivalents in Serbian points to the fact that probably lexical concepts are in question that have not yet been included in SWN, and that their addition to this wordnet should thus be considered. However, in the case of absence of translational equivalents in English, one should bear in mind that at this moment the morphological expansion is available for Serbian only. It is thus understandable why the English form ‘games’ which is the equivalent for the Serbian form ‘igre’ was not recognized in the first segment, marked as n4.

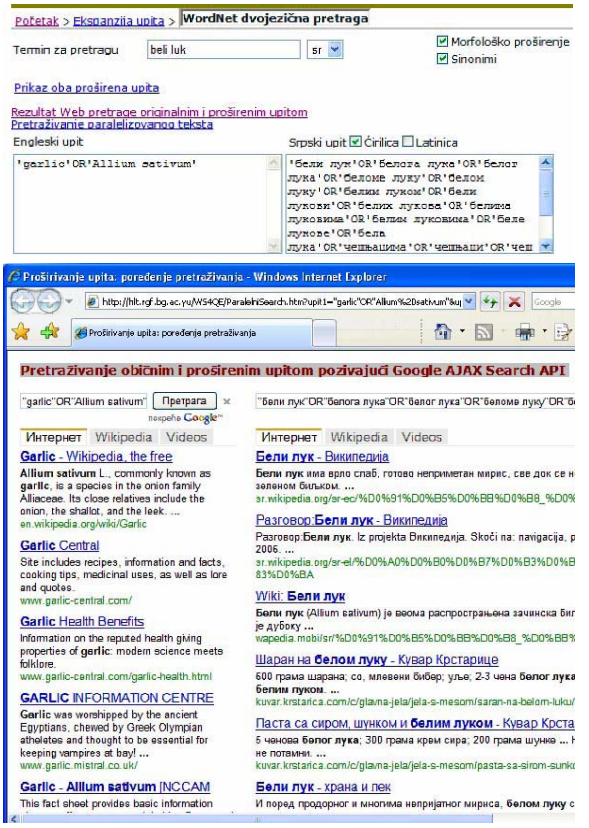

Figure 12. Bilingual query expansion and results of Google search

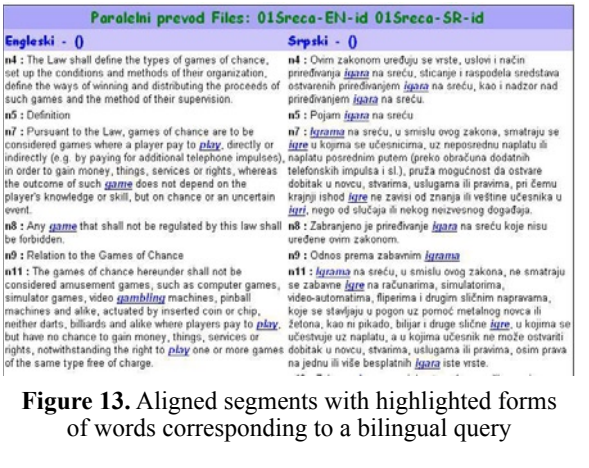

Figure 13. Aligned segments with highlighted forms of words corresponding to a bilingual query

5 Conclusion

The work on the integration of resources for Serbian that started several years ago will be continued in several directions. In the first place, this pertains to further development of the module enabling searches with all morphological forms of compounds. Since this search, besides the dictionary of compounds, relies also on specific rules for the generation of their inflectional 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.

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 of some WS4LR functions, with the possibility of access to the WS4QE web application is also being considered.

[1]On finite transducers see e.g. Mohri, 1997.
[2]http://igm.univ-mlv.fr/~unitex/
[3]http://www.nooj4nlp.net
[4]Let us remind that "cx" represents the letter ć in the Aurora code.
[5]For details on TMX format see http://www.lisa.org/tmx/tmx.htm
[6]The version WS4QE under development is available at http://hlt.rgf.bg.ac.yu/WS4QE/

6 References

Barzilay, R., McKeown, K. R. (2001) "Extracting paraphrases from a parallel corpus", *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France 2001, pp. 50 – 57.

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

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

Fellbaum, C. (ed.). (1998): *WordNet: An Electronic Lexical Database*, Cambridge, Mass. MIT Press.

Francopoulo G., Bel N., George M., Calzolari N., Monachini M., Pet M., Soria C. (2006) "Lexical Markup Framework (LMF) for NLP Multilingual Resources", *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, Australia, pp. 1–8.

Gale, W., Church, K. (1993) "A Program for Aligning Sentences in Bilingual Corpora", *Computational linguistics* 19(1), pp. 75-102.

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

Krstev C., Vitas D., Pavlović-Lažetić, G. (2008) "Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian", in *Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003*, Zybatow, Gerhild et al. (eds.), Peter Lang: Frankfurt am Main, pp. 3-17.

Mohri, M. (1997) "Finite-state transducers in language and speech processing", *Computational Linguistics*, vol. 23 , no. 2, pp. 269 – 311.

Ohmori K., Higashida M. (1999) "Extracting bilingual collocations from non-aligned parallel corpora", *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, Chester, England; pp. 88-97.

Silberztein, M. (1993) *Le dictionnaire électronique et analyse automatique de textes: Le systeme INTEX*, Paris, Masson.

Steinberger, R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufiş D., Varga D. (2006) "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages", *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, pp. 2142-2147.

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

Vitas, D., Krstev, C. (2005) "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.

Vossen, P., (ed.) (1998) *EuroWordNet. A multilingual database with lexical semantic network,* Dordrecht, Kluwer.