The Nooj System as Module within an Integrated Language Processing Environment

Ranka Stanković, Duško Vitas, Cvetana Krstev

The NooJ system as module within an integrated language processing environment

Ranka Stanković, ranka@rgf.bg.ac.yu
Duško Vitas, vitas@matf.bg.ac.yu
Cvetana Krstev, cvetena@matf.bg.ac.yu

1. Introduction

In this paper we describe the main structure and possible applications of one integrated environment for linguistic research that contains NooJ as one of its main modules. This environment named WS4LR (WorkStation for Lexical Resources) has been developed within the Human Language Technology Group (HLT) at the Faculty of Mathematics, University of Belgrade, and is aimed at manipulating heterogeneous lexical resources developed in the course of many years and within different projects. The tool handles morphological dictionaries, wordnets, aligned texts and transducers and has already proved very useful for various tasks. Although it has so far been used mainly for Serbian, WS4LR is not language dependent and can be successfully used for resources in other languages provided that they follow the described formats and methodologies.

The integration of NooJ with other language resources was aimed in the first place at integrating the morphological power of NooJ with the semantic and multilingual power of wordnets. This integration enables the performance of many sophisticated tasks such as query expansion. By this we mean the techniques in which a query, serving as input to a document retriever, is transformed in some way in order to improve the performance of document retrieval. Namely, a search by a concept instead of a search by a single word form is recognized as a very important new direction in information retrieval and related areas. If query is further combined with ILI, a multilingual wordnet pivot, the possibility of searching text resources (web, corpus, text) in different languages with a single query is opened. NooJ supports morphological query expansion and expansion of queries by graphs and regular expressions. The integration of the morphological power of NooJ with the semantic and multilingual power of wordnets may best be illustrated by concordance production using various search requirements such as extraction by simple strings, lemmas (with all their inflectional forms) or concepts (all or some lemmas from chosen synsets and/or their hypernyms).

The developed framework also makes possible the federation of NooJ resources with Prolex, multilingual dictionary of proper names which is based on one ontology built around the conceptual proper name and its relations. This adds additional functionality for information retrieval, indexing, machine aided translation, machine translation, and alignment of multilingual texts.

WS4LR handles aligned texts as well. A pair of semantically equivalent texts in different languages, such as an original text and its translation, that are aligned on a structural level (paragraph, sentence, phrase, etc.) is known as an aligned text or bitext. One of the supported formats is the Translation Memory eXchange format (TMX, TMX Specification 2005), a standard format for representing aligned texts. The integrated functions and resources enable queries to be posed in one language and bitext to be searched in the same or other language.

For example, if a query consists of Serbian word '*kompjuter'*, it can be expanded by Serbian wordnet to *'računar, kompjuter*', and then transformed by ILI (interlingual index) to query consisting of a set of English keywords: '*computer, computing machine, computing device, data processor, electronic computer, information processing system'*. Similarly, English query '*document*', is expanded by English wordnet to *'document, written document, papers'* which is then transformed by ILI to a set of Serbian keywords: '*dokument*, *papir*, *akt'* that can be further expanded to all inflectional form by NooJ inflectional graphs: '*document, dokumenta, dokumentu, dokumentom,..'*

2. Integrated environment for linguistic research

2.1. Motivation

The Human Language Technology group has been developing a variety of lexical resources over a long period, reaching a considerable volume to date. These resources have been developed for many years, and it is understandable that they have been conceived within different projects and frameworks, both from the conceptual and the technological perspective. Despite that the HLT group made every reasonable effort to keep the constantly growing pool of resources as coherent and standardized as possible, a certain level of heterogeneity was inescapable. Henceforth, due to the growing of the volume of resources as well as their heterogeneity, there was a rising need for development of a tool that would facilitate the maintenance, exploitation and integration of available resources as well as their further development. Embarking on this task, the HLT group produced

an integrated and easily adjustable tool, the workstation for language resources, labeled WS4LR, which greatly enhances the potentials of manipulating each particular resource as well as several resources simultaneously. Exploiting the synergy of various resources, this tool proved very useful in many HLT tasks, including wordnet development.

2.2. Structure

WS4LR is made up of several modules which perform the following main functions (Figure 1):

- management of a system of electronic dictionaries which consists of morphological dictionaries of lemmas for simple and compound words but also of bilingual and multilingual dictionaries
- development and refinement of wordnets, with simultaneous usage of wordnets for different languages
- conversions from different formats such as one character encoding set to another or from one resource format to another
- manipulation of bilingual aligned texts, allowing for various forms of their presentation and usage

This software tool is developed in C# and operates on the .NET platform. An important feature of WS4LR is its flexibility expressed both by the possibility of setting environment parameters and by the possibility of invoking command-line routines and using external Perl, Awk, and XSLT scripts.

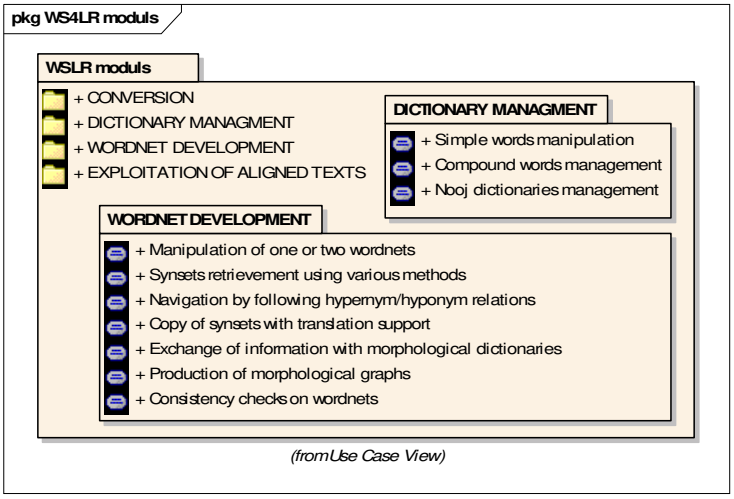

Figure 1. The UML diagram with WS4LR modules

2.3. Conversion

The conversion module is crucial since it adds to the flexibility of resources exploitation. Conversion from one character encoding set to another is very important for languages such as Serbian, where two alphabets, Cyrillic and Latin are equally used. WS4LR enables the exploitation of language resources both in Cyrillic and Latin alphabet, as well as in a special encoding, that uses the ASCII character set and that can be unambiguously transformed into Serbian Latin or Serbian Cyrillic alphabet. In that special encoding, for example, "sx" is used as a code for "š". With the emergence of NooJ the available resources have been expanded by a conversion of all Intex DELAS dictionaries into NooJ dictionaries, in the original, ASCII format as well as in the Latin and Cyrillic version.

WS4LR offers to the user the option to apply the transformation only to a part of the file, such as an XML file where only the text should be converted while the XML tags should not be altered. Similarly, when a dictionary type file is transformed, lemmas and word forms are converted, while part of speech and grammatical codes are not. The user can choose a conversion Perl or awk script suitable for the specific file type, or produce his/her own script easily.

ISO standards provide a common model for the creation and use of lexical resources, manage the exchange of data between and among these resources, and to enable the merging of large numbers of different individual electronic resources to form large global electronic resources, so conversion of NooJ resources to LMF format (Lexical markup framework) (ISO LMF 2006) is also included in this environment.

3. Lexical resources management

3.1. Dictionary Management

This module enables concurrent manipulation of a set of dictionaries of lemmas, simple words or compounds, distributed in several files. The organization of dictionaries in separate files is important from the practical point of view since smaller files are easier to manipulate. An even more important reason is the fact that in text recognition by NooJ the usage of all dictionaries is not always necessary, or even recommended.

Morphological dictionaries are of great importance for highly inflective languages, such as Slavic languages. The absence of morphological information in wordnets has turned out to be a serious flaw in many applications. Thus the possibility, offered by WS4LR to simultaneously exploit both resources proved to be a great advantage in wordnet development. Given the importance of morphological, but also bilingual

and multilingual dictionaries in wordnet development, we will now briefly describe the basic features of the dictionary management module.

The lemma in a morphological dictionary of simple words has the following format: *lemma.Knnn [+SinSem]*,* where *lemma* is the word form usually used in traditional dictionaries, *K* represents the part of speech (noun, verb, adjective, etc.), and *nnn* the inflectional class code of the lemma, whose characteristics are described by a corresponding transductor labeled *Knnn.* A set of optional tags *+SinSem* follows, which describe the syntactic, semantic, derivational and other properties of the lemma (Courtois & Silberztein 1990).

The WS4LR module has been developed to enable the entry, editing and review of lemmas of simple and compound words. Dictionaries are organized in modular fashion, in several sub-dictionaries as separate files. Without going into details of dictionary management, we will just point out that the dictionary management module enables the user to modify or delete all the information attached to a lemma, or the lemma itself, as well as to add new entries (Figure 2). A new entry can be generated from scratch or by copying and then editing an existing lemma, which in some cases facilitates the work. The regular expression or a FST graph describing the inflectional properties of the selected lemma can be inspected and corrected if found inadequate. The FST can also be applied to a lemma, and the user can inspect the produsedforms in order to see whether the chosen FST is appropriate to the lemma.

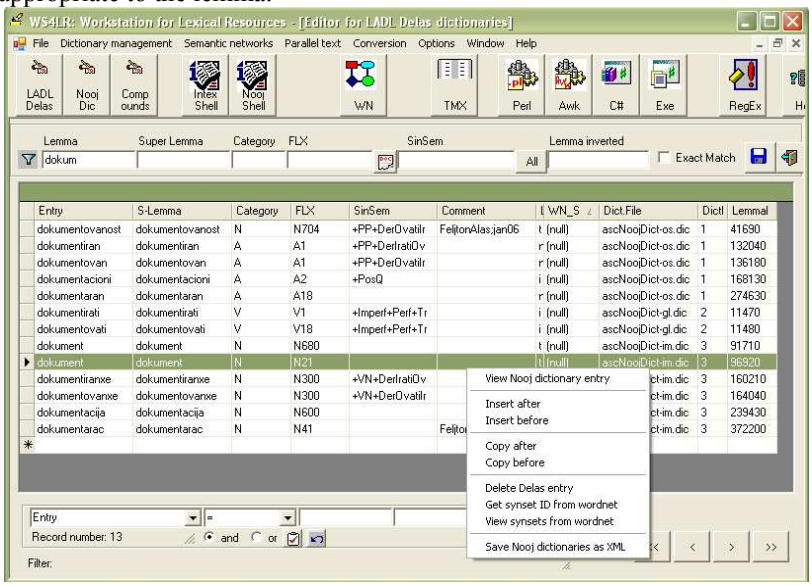

Figure 2. Retrieval of all dictionary entries starting with “dokum”

An important feature of this module is the ability of retrieving efficiently a subset of lemmas by matching the lemmas, their part of speech, inflectional class code, syntactic and semantic markers or their Boolean combination. For instance, one can look for all the dictionary entries starting or ending with a search string. The latter is particularly useful when the inflectional code of a new lemma is being established, since this code depends often on the lemma ending.

The form for viewing and/or editing a single entry is presented in Figure 3. The format of the lemmas for compound words in NooJ is more complex, but it basically relies on the same principles.

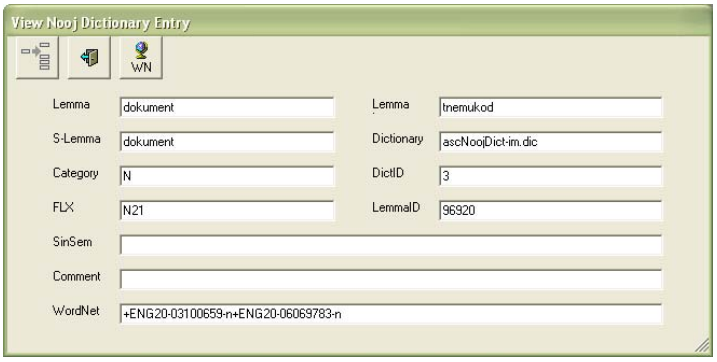

Figure 3. Form for viewing a NooJ dictionary entry

As we have already mentioned, WS4LR handles besides bilingual word lists also multilingual dictionaries, such as Prolex, the multilingual dictionary of proper names based on an ontology built around the conceptual proper name and its relations (Krstev 2005). This adds additional functionality to the integration of lexical resources offered by WS4LR in various tasks.

3.2. Wordnet Manipulation

The wordnet management module supports search of wordnets, stored in XML format, their visualization, as well as their development and refinement. When this module is activated, the main form opens with two wordnet windows, thus offering to the user the possibility to work with one or two wordnets. If the user decides to work with two wordnets in parallel, he/she can always synchronize them via the ILI. The equivalent synsets in different languages are linked to the same Inter-Lingual Index (ILI) thus connecting monolingual wordnets in a global lexical-semantic network (Vossen 1998). The main form for wordnet management also opens a window with a bilingual word list (Figure 4).

Synsets can be retrieved from wordnets into the two available wordnet windows using various methods, from simple string matching to complex

Xpath expressions. The user can, for example, specify one or two strings, depending on whether he/she wants to search one or both wordnets for synsets containing words that match the string(s). The user can also specify whether he/she wants an exact match or not, and in the latter case the system will retrieve not only all synsets with words matching the search string(s), but also those that contain words that have as substrings the specified string(s).

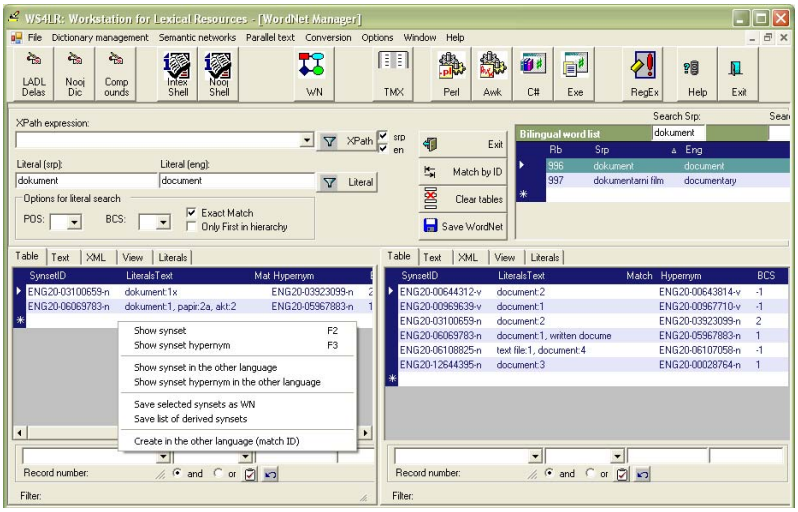

Figure 4. The main form for wordnet management

The user can also use an Xpath expression to retrieve synsets on basis of various other criteria, such as the domain synsets belong to. Thus, for instance, with the expression: "*//SYNSET[DOMAIN='administration']*" the user can retrieve all synsets from the wordnet that belong to the domain of administration, or more precisely, that contain the XML tag <DOMAIN> with the content "*administration*". WS4LR offers predefined Xpath expressions, but the user can easily define his/her own expressions.

Once the user has retrieved the synsets of interest from the wordnet, he/she can now proceed to their modification or generation of new synsets. Every retrieved synset can be visualized in various forms of visualization: as text, XML or hypernym/hyponym tree (Figure 5). There is also an edit view for the synset which allows the user to modify the synset contents: word-sense pairs, definition, usage, but also other properties, such as semantic relations to other synsets. In the hypernym/hyponym view, the user can easily navigate through its hypernym/hyponym tree and proceed to further modifications of synsets.

WS4LR allows for adding of new synsets to wordnets using predefined forms. The two main problems is how to place the synset in the conceptual network, and how to select the appropriate word-sense pairs to represent the

concept of the synset. The module enables easy production of Intex type graphs that would locate all literals from a chosen synset in a text, with or without synset hypernyms. User can than use the generated graphs in Intex environment to perform the actual retrieval, or import them in NooJ and convert to syntactic grammars in order to perform the same task.

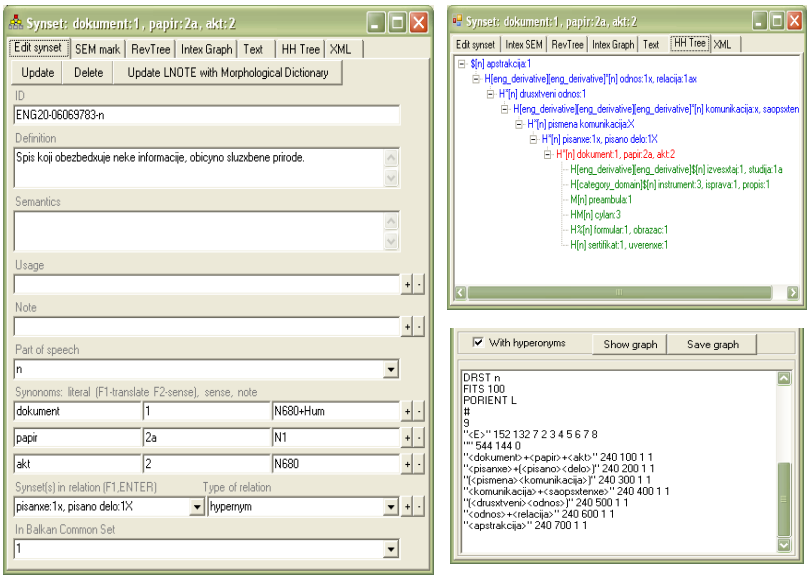

Figure 5. The edit view, the hypernym/hyponym and graph view of a synset

3.3. The exchange of information

3.3.1. Usually, the only grammatical information accompanying the synset literal in a wordnet is the PoS, and it has to be the same for all literals in one synset. Enriching the synsets with information from the morphological dictionary makes the usage of a wordnet in an information retrieval task more efficient. In a number of cases this additional information disambiguates the otherwise homonymous literals. This additional morphological, syntactic, and semantic information can be transferred from the morphological dictionary and associated to each synset literal in wordnet. For instance, in the following two synsets

(obaviti:A1x, uraditi:4) ↔ (do:3, perform:4)
(okruzxiti:4, obaviti:B1v) ↔ (smother:1, surround:3)

the homographous literal *obaviti* appears. In both cases its is a verb but with two different inflectional paradigms (for the verb in the first synset the first person singular present form is *obavim* while for the verb in the second synset it is *obavijem*). That is, the inflectional paradigm of a verb, and same stands for other parts-of speech, cannot be automatically determined.

Information about the inflectional properties can be attached to each literal string in wordnet, in a form of a XML element included into the literal element. For this kind of information the element <LNOTE> can be used that is contained in the element <LITERAL> in the core XML schema of the Balkanet project (Tufiş 2004).

3.3.2. The information from wordnet can be successfully used to enrich the morphological dictionaries, namely the wordnet hierarchy can be used to add semantic information to word entries in morphological dictionaries. Some basic semantic information has already been attached to simple word entries, such as +Hum (human) and +Bot (botanic) for nouns, and +Col (colour) and +Mat (material) for adjectives. The use of wordnet enables a more systematic and more detailed attachment of such marks. Moreover, the attachment can be modeled according to the envisaged application. The hierarchy corresponding to the following branch in PWN (Prinston wordnet, Fellbaum 1998)

abstraction:6
attribute:2
property:3
sound property:1
sound:1

can yield the addition of appropriate semantic marks to the entry

glas 'voice' (hyponym of 'sound:1'):
glas,N16+Snd+SndProp+Prop+Attr+Abstr.

Depending on the application the depth of the tree hierarchy and/or its level can be chosen.

Since only some basic semantic information has been incorporated in the Serbian morphological dictionary, there are a number of identical entries that is, apparently same lemmas with identical inflectional classes and morphosyntactic information attached to them, but actually representing different lemmas that can not be distinguished. That is the case, for instance, with the double entry *cyelo,N300* that represents both *(brow:1, forehead:1)* and *(cello:1, violoncello:1)*. By adding the information obtained from the wordnet hypernym/hyponym relations these two entries can be distinguished: for instance,

cyelo,N300+BodyPart and
cyelo,N300+Artifact,

or

cyelo,N300+Thing+BodyPart+Feature and
cyelo,N300+Artifact+Device+MusicInstr

if more semantics is used.

The principal feature of WS4LR is its capability to work with a wordnet and morphological dictionaries in parallel and to enable transfer of

information from one type of a resource to the other. In order to perform these tasks special tab-pages of the edit form are designed.

For instance, the tab-page "Update with morphological dictionary" enables the inclusion of morphosyntactic information from Serbian morphological dictionary into a working synset. An element LNOTE, which is in the content of the LITERAL element in the XSD schema common to all Balkan languages is used in the Serbian wordnet for morphosyntactic information specific for this literal. This information is automatically retrieved from the morphological dictionaries. If more than one instance is retrieved from these dictionaries, the user can choose the appropriate one. Moreover, he can modify (delete or add) the automatically retrieved information.

The second tab-page "SEM mark" (Figure 6) is used for retrieving semantic information that is going to become the content of SEM element, from all the working synset's hypernyms up to the root element. SEM element, the newly introduced element in a wordnet's XML schema, is a semantic tag, that points to a specific sense of that synset in the wordnet database. Then all the lemmas corresponding to working synset's literals are retrieved from the morphological dictionary, and a string of plus sign marks is formed, which the user can choose to add to the retrieved lemmas. If due to the addition of semantic information one lemma has to be separated in two or more lemmas, the copies of the original lemma can be made and appropriate semantic information added to each of them.

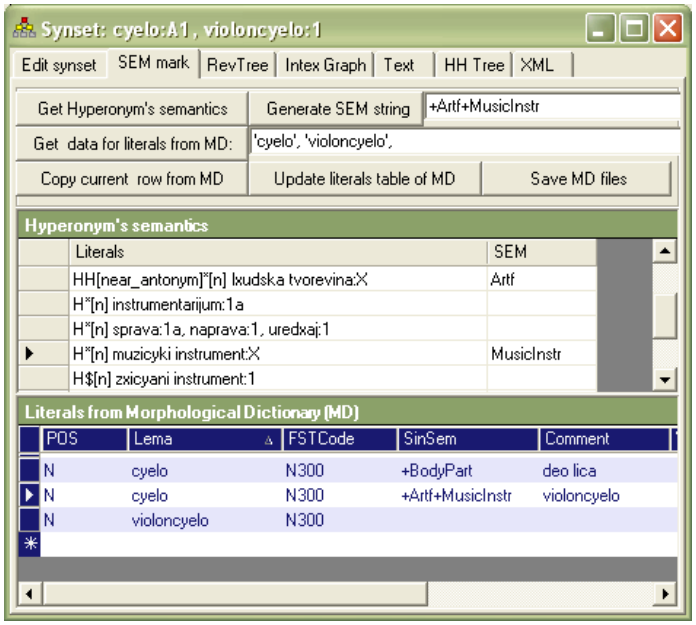

Figure 6. The semantic separation of the lemma *cyelo*

4. Textual resources management

4.1. Parallel Text Management

The WS4LR module for management of aligned parallel texts uses texts which have previously been aligned using Xalign as an alignment tool (Bonhomme 2001). Parallel texts which usually originate from a text in one language and its translation in another, are often aligned at a certain level (paragraph, sentence, etc) by matching the corresponding segments of the original and its translation.

The module converts these texts to the Translation Memory eXchange (TMX) format, which is becoming the standard format for aligned texts. Figure 7 depicts the form with different possibilities for TMX document management. Aligned texts can be visualized in various ways by choosing the appropriate XSLT stylesheet. Namely, the user can obtain the aligned text in HTML format, but also in textual, XML, tabular or TMX format.

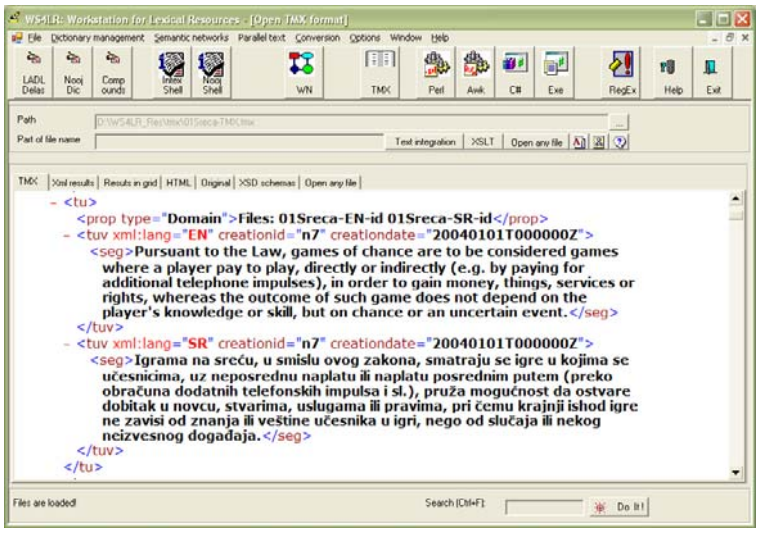

Figure 7. The part of a TMX document

4.2. Query Expansion

Query expansion is a name given to a class of techniques in which a query serving as input to a document retriever is evolved in some way with the intent to improve the document retriever's performance, according to some metric. Query expansion is particularly applicable to document retrieval components that provide a Boolean query model, because of the expressiveness of the syntax and ease of modifying existing queries. Query

expansion is a technique used to boost performance of a document retrieval engine. Common methods of query expansion for Boolean keyword-based document retrieval engines include inserting query terms, such as alternate inflectional or derivational forms generated from existing query terms, or dropping query terms that are, for example, deemed to be too restrictive (Bilotti 2004).

This section introduces different ways to expand a query: morphologically, semantically, in another language. Serbian language has a rich morphology, so simple string matching can rarely fulfill user demands. Also, the possibility of semantic extension, namely, a search by concept instead of by single word form is recognized as a very important direction in information retrieval and related areas. Combined with the wordnet ILI, this approach opens the possibility of searching text resources (web, corpus, text) in different languages with a single query.

Powerful linguistic tools such as NooJ, though inherently multilingual since resources for it have been developed for many languages, at present do not support simultaneous work with different languages. With WS4LR we have tried to, at least partially, overcome this shortcoming and enable better exploitation of aligned texts as resources of great value. This is achieved by integration of all the resources supported by WS4LR. It may best be illustrated by concordance production using various search criteria such as simple strings, lemmas (with all their inflectional forms) or concepts (all or some literals from the chosen synsets, and/or their hypernyms, and/or synsets connected to the chosen synset(s) by some other semantic relation supported by the underlying wordnet).

The result of user interaction with WS4LR is recorded formally in a XML document. The query specification includes, apart from the lemma, the alphabets used, followed by morphological, semantic and multilingual expansion information. Within the query, the file with textual resources for application of the expanded query is also specified, namely the file which is going to be searched, as well as the resulting files which can be in different formats.

Figure 8 presents a form with query expansion for lemma: *dokument*, with morphological dictionaries and transducers as the morphological resource and Serbian wordnet as the resource for semantic expansion. The user can also use the translation equivalence option which is aimed at locating equivalences in target language for occurrences found in the source language. This is done on the basis of data from wordnets for corresponding languages. Thus the search criteria including *dokument, papir, akt* is expanded with *document, written document, papers*. Optionally, the user can include literals from hypernyms of the selected synsets, so the search criteria includes *pisanxe, pisano delo* in Serbian and *representation, writing, written material, piece of writing* in English part of the query.

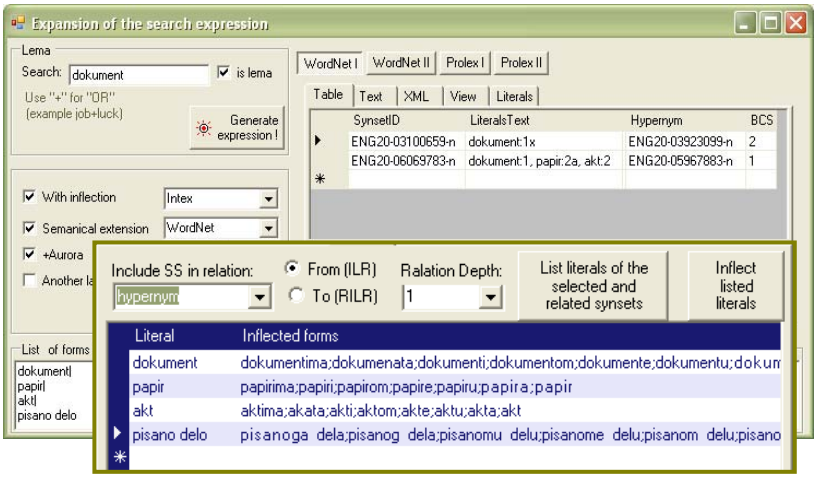

Figure 8. The form for the expansion of the search based on a single keyword

In aligned segments retrieved on the basis of the generated query, occurrences that satisfy a query in the source and target languages are highlighted (Figure 9). A filtered TMX document is transformed into XML, TXT and HTML output files by XSLT transformations.

Paralelni prevod Files: 02Porez-EN-id 02Porez-SR-id

Engleski -EN (dokument)	Srpski -SR (dokument)
n206 : Tax may also be paid by the purchase of a prescribed security instrument (payment stamps, supplemental postal stamps, fiscal excise stamps etc.), in the cases stipulated by the law.	**n205 :** Porez se može platiti i kupovinom vrednosnog ***papira*** (taksene marke, doplatne poštanske marke, fiskalne akcizne markice i sl.) u slučajevima propisanim zakonom.
n218 : 4. the security instrument indicated in Article 67, paragraph 3 of this Law was properly cancelled or purchased.	**n217 :** 4) na propisani način poništen, odnosno kupljen vrednosni ***papir*** iz člana 67. stav 3. ovog zakona.
n220 : The day of the settlement of tax liability by compensation shall be considered the day when the Tax Administration has received the ***document*** on performed compensation, deposited with the payment operations agent and signed by all parties involved and legally stamped.	**n219 :** Danom namirenja poreske obaveze putem kompenzacije smatra se dan kada je Poreskoj upravi dostavljen, od strane svih učesnika u kompenzaciji potpisan i overen ***dokument*** o izvršenoj kompenzaciji koji je realizovan kod nosioca platnog prometa.
n222 : The day of the settlement of tax liability by conversion of the tax claim into permanent shares of the Republic in the taxpayer's capital shall be considered the day when the Government adopted the act on conversion.	**n221 :** Danom namirenja poreske obaveze putem konverzije poreskog potraživanja u trajni ulog Republike u kapitalu poreskog obveznika smatra se dan kada je Vlada donela ***akt*** o konverziji.

Figure 9. An excerpt from a aligned text that shows segments that satisfy the expanded query

However, a disjunction of several lemmas in the initial query is also suported, where the lemmas are connected by a „+“ sign. The expanded query will consist of a union of all separate expansions for each lemma. For example, Figure 10 presents a form with query expansion for the disjunction of two lemmas: *dokument+zakon* (“+” stands for “OR”).

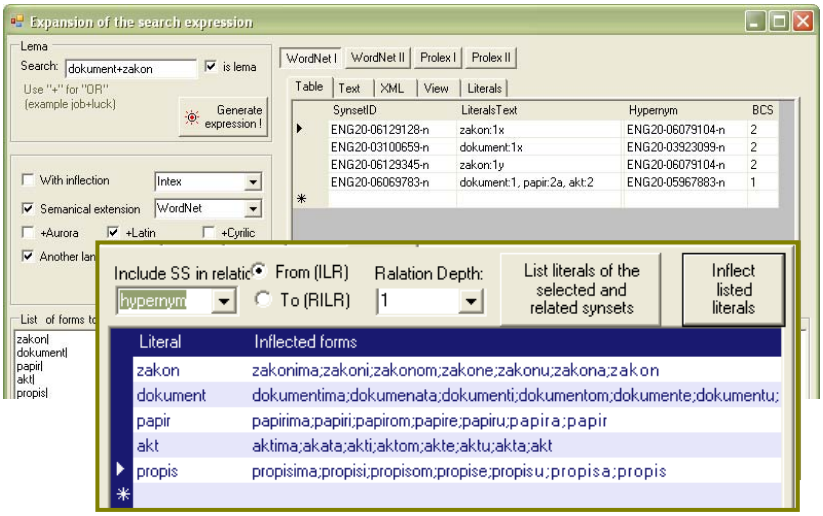

Figure 10. The form for the expansion of the search based on a disjunction of two keywords

In aligned segments retrieved on the basis of the generated query, occurrences that correspond to the expanded search in the source and target languages are highlighted (Figure 11).

n52 : Disabled persons may establish employment relations under the conditions and in the manner provided by this ***Law***, unless otherwise provided by a special ***law***.	**n52 :** Invalidna lica zasnivaju radni odnos pod uslovima i na način utvrđen ovim ***zakonom***, ako posebnim ***zakonom*** nije drukčije određeno.
n56 : In case the employee fails to ***act*** in accordance with Paragraph 1 of this Article, the employer may cancel the employee's labour contract.	**n56 :** U slučaju da zaposleni ne postupi u skladu sa stavom 1. ovog člana, poslodavac može zaposlenom da otkaže ugovor o radu.
n58 : Foreign nationals or persons without citizenship may establish employment relations under the conditions provided by this ***Law*** and a special ***law***.	**n58 :** Strani državljanin ili lice bez državljanstva može da zasnuje radni odnos pod uslovima utvrđenim ovim ***zakonom*** i posebnim ***zakonom***.
n63 : When establishing the employment relations the employee shall provide the employer with the ***documents*** proving that the former meets the conditions for working.	**n63 :** Zaposleni je dužan da, prilikom zasnivanja radnog odnosa, poslodavcu dostavi ***dokumenta*** kojima se dokazuje ispunjenost uslova za rad.
n141 : The employer may establish employment relations with a person who is entering employment for the first time, in the capacity of a trainee, for an occupation for which such a person has acquired appropriate education, if that is prescribed by ***law*** or by the ***enactment*** referred to in Article 13 of this ***Law*** as a condition for doing a particular job.	**n141 :** Poslodavac može da zasnuje radni odnos sa licem koje prvi put zasniva radni odnos, u svojstvu pripravnika, za zanimanje za koje je to lice steklo određenu školsku spremu, ako je to kao uslov za rad na određenim poslovima utvrđeno ***zakonom*** ili ***aktom*** iz člana 13. ovog ***zakona***.

Figure 11. Retrieved aligned segments with highlighted occurrences that correspond to search by disjunction *document+zakon*, expanded both morphologically and semantically

4.3. Implementing the Query with NooJ

Once the query has been expanded using various WS4LR modules it can be submitted to NooJ for implementation. To that end the query has to be in the form of a graph or a regular expression. While regular expressions can be dynamically generated, graphs must be prepared through the NooJ interface. Once the graph or regular expression is ready, NooJ offers two possibilities for their application to a text: the dynamic library NoojEngine.dll and a command-line program: noojapply.exe.

In its Community edition, NooJ's functions are available via a .NET dynamic library, *noojengine.dll*, constituted by a set of public object classes and methods. These classes and methods can be used by any .NET application, so *noojengine.dll* is linked with WS4LR and available NooJ functions are used. Another, at the moment, more powerful option is to use the *noojapply.exe* program, available in Standard edition of NooJ. Noojapply.exe allows users to apply dictionaries and grammars automatically to texts or corpora. This paper presents two search examples using NooJ regular expressions and the NooJ syntactic grammars. Different options in application of *noojapply.exe* are presented on the left side in Figure 12.

First usage of noojapply.exe offered in the list is the compilation of dictionary, while the other four produce concordances in text, XML and table output by application of different lexical resources. After selecting the type of *noojapply.exe* usage the user can choose the dictionaries and morphological grammars that he wishes to apply from a list of available lexical resources. Next, one or more text files or corpus should be selected from a list and, finally, syntactic resources: syntactic graphs or regular expressions.

Depending on the selected *noojapply.exe* option the resource selection process differs slightly. For example, if second option is selected ("Apply lex-resources to texts") then syntactic resources should not be chosen, and if the last option is on ("Apply query to corpus"), then the user selects only a query and a corpus.

Figure 12 presents results in the form of concordances for the query: *kompjuter,* which was automatically expanded with its synonyms by a correspoding wordnet synset and transformed to NooJ regular expression <*racynar*>+<*kompjuter*>. Expanded query is saved in "nox" type of syntactic resource.

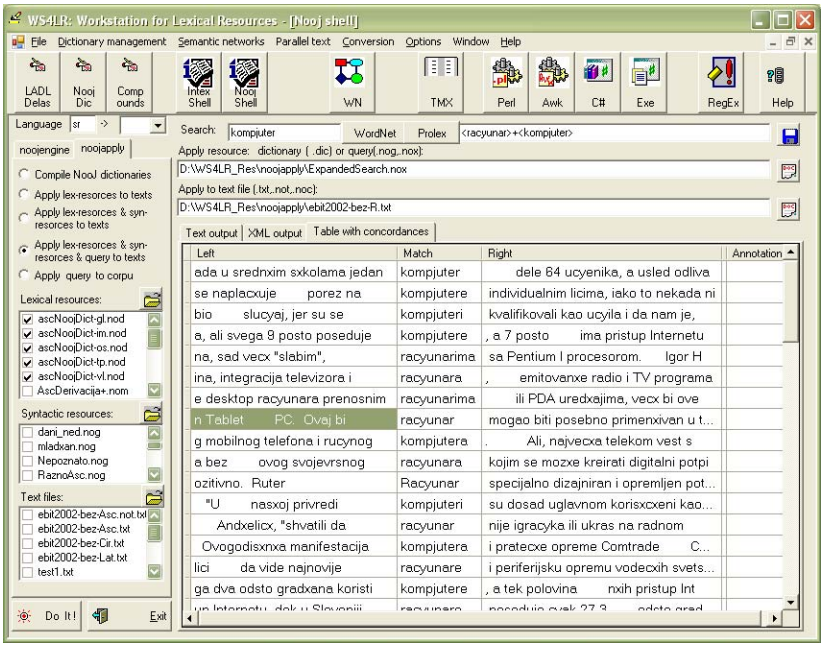

Figure 12. An example of the search with a NooJ regular expression

With regular expressions user can use more general patterns since he/she is not limited to queries using just one or more lemmas. Namely, the query can have a more general form, such as: *<A+PosQ> <oficir>*, where *<A+PosQ> represents all relational* adjectives. This query can be automatically expanded with wordnet by adding the hyponyms of *oficir* and transformed to NooJ regular expression <A+PosQ>(<oficir>+<kapetan>). Figure 13 presents concordances for the specified query, which recognizes word sequences with relational adjectives followed by *oficir* 'officer' or *kapetan* 'captain'.

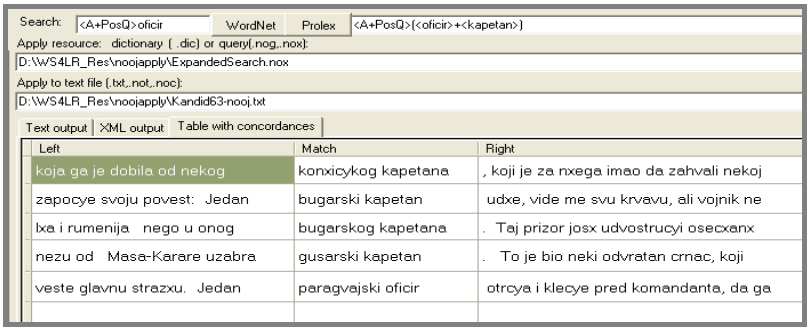

Figure 13. An example of the search with a NooJ regular expression expanded by a wordnet

5. Conclusions

Inflection for the target language in aligned texts is not yet supported. Namely, the translation equivalence option finds all synsets that contain the literals corresponding to the searching lemma in the wordnet of a source language and then the corresponding synsets in the target wordnet via the ILI. The search in the target language is then performed with synset literals only, without their inflected forms. We plan to include these forms in the search as well.

The power of integrated language resources tools such as WS4LR has been illustrated by its different function. It has also been shown how WS4LR can be applied to query expansion. Without such a tool the results of a query specified by a user in the search of a specific concept would be reduced to only a fraction of the results obtained by query expansion. Although WS4LR has been used mainly for Serbian language resources, it is by no means language dependent. The only prerequisite is that the resources exist or are being developed according to the described formats and methodologies. Of course, not all of the resources need to exist. The user can work only on the resources he/she develops or that are available to him/her and modules that support them.

The development of WS4LR will continue as we intend to incorporate in it some more sophisticated features. A web service is already being developed with query expansion functionality. The integration with MS Word would be very useful for numerous users, and to that end we plan to incorporate some of WS4LR functions into MS Word as well.

6. References

Bilotti, M., Query Expansion Techniques for Question Answering, Massachusetts Institute of Technology (2004)

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

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

Fellbaum, C. (ed.) "WordNet: An Electronic Lexical Database", The MIT Press (1998)

ISO LMF, Language resource management - Lexical markup framework (LMF), (2006), ISO/TC 37/SC 4 N130 Rev.9, ISO CD 24613:2006.

Krstev C., Pavlović-Lažetić G., Vitas D., Obradović I.: Using Textual and Lexical Resources in Developing Serbian Wordnet. Romanian Journal of Information Science and Technology, Romanian Academy,

Publishing House of the Romanian Academy, vol. 7, No. 1-2, pp. 147-161, (2004)

Krstev C., Vitas D., Stanković R., Obradović I., Pavlović-Lažetić G.: Combining Heterogeneous Lexical Resources. Proc. of the Fourth International Conference LREC, Lisbon, Portugal, May 2004, vol. 4, pp. 1103-1106 (2004)

Krstev C., Vitas D., Maurel D., Tran M. (2005), Multilingual Ontology of Proper Names, *Second Language & Technology Conference*, 116-119, Poznan, Poland, 21-23 avril.

Krstev C., Stanković R., Vitas D., Obradović I.: WS4LR: A Workstation for Lexical Resources. Proc. of the Fifth International Conference LREC, Genova, Italy, May 2006

Silberztein, M. NooJ Manual, Université de Franche-Comté (2007)

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

TMX 1.4b Specifacion, OSCAR Recommendation, 26 april 2005, http://www.lisa.org/tmx/

Vitas D., Krstev C., Obradović I., Popović Lj., Pavlović-Lažetić G.: Processing Serbian Written Texts: An Overview of Resources and Basic Tools., Workshop on Balkan Language Resources and Tools, Thessaloniki, Greece, eds, S. Piperidis and V. Karkaletsis, pp. 97-104, 2003.

Vossen, P. (ed.): *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic publishers, Dirdrecht, 1998.