An Approach to Development of Bilingual Lexical Resources

Stanković Ranka, Obradović Ivan, Trtovac Aleksandra

An Approach to Development of Bilingual Lexical Resources

Ranka Stanković
University of Belgrade
Faculty of Mining and Geology
Đušina 7, Belgrade
Serbia
+381 11 3219 148
ranka@rgf.bg.ac.rs

Ivan Obradović
University of Belgrade
Faculty of Mining and Geology
Đušina 7, Belgrade
Serbia
+381 11 3219 259
ivano@rgf.bg.ac.rs

Aleksandra Trtovac
University of Belgrade
University Library "Svetozar Marković"
Bulevar kralja Aleksandra 71, Belgrade
Serbia
+381 11 3370 211
aleksandra@unilib.bg.ac.rs

ABSTRACT
This paper outlines how Bibliša, a tool initially designed for search of digital libraries of articles from bilingual e-journals in the form of TMX documents, is used for development of a new bilingual lexical resource. The approach relies on already available resources, Serbian morphological e-dictionaries, Serbian and English wordnets connected via the interlingual index, and a bilingual Dictionary of Librarianship, as well as on a TMX document collection generated from aligned Serbian-English journal articles published in INFOtheca, a scientific journal in the area of Library and Information Sciences. The aim of the new resource, Biblimir, is to help overcome the shortcomings of existing bilingual resources and hence improve the performance of Bibliša.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection*

General Terms
Documentation, Languages

Keywords
Digital libraries, aligned parallel texts, TMX document collections, multilingual lexical resources, bilingual search

1. INTRODUCTION
Multilingual information exchange is growing in importance and several current large scale European projects are tackling this issue from different perspectives. One of them, META-NET, is aimed at building technological foundations of a multilingual European information society, through a shared vision and strategic research agenda. Among its core objectives is META-SHARE, an open distributed facility for sharing and exchange of language resources among various European languages [Piperidis, 2012].

Another project, the Multilingual Web Initiative, led by W3C, is a thematic network, exploring standards and best practices supporting the creation, localization and use of multilingual web-based information [Filip et al., 2012]. Hence, the importance of multilingual language resources is rapidly increasing, as is their availability on the web, followed by a need for efficient methods and strategies for developing and searching these resources.

BCI'12, September 16–20, 2012, Novi Sad, Serbia.

Local Proceedings also appeared in ISBN 978-86-7031-200-5, Faculty of Sciences, University of Novi Sad.

Multilingual textual repositories, such as digital libraries of e-journals represent a specific type of language resources. Efficient search of these resources usually relies on specific language tools, which often use other available resources, such as e-dictionaries, wordnets and the like. An interesting tool for keyword extraction in multilingual textual repositories is described in [Pammzi et al., 2006]. The tool extracts keyword collocations from documents and generates multiword keywords. The paper also outlines linguistic criteria used for building language resources for French, Italian, and German, and the use of multi-term descriptors as a means to better identify the content.

The Human Language Technology group at the University of Belgrade developed Bibliša (http://hlt.rgf.bg.ac.rs/Biblisha), a tool aimed at enhancement of search possibilities in digital libraries of e-journals, or more precisely, textual resources representing collections of TMX documents with corresponding metadata. For testing and evaluation of Bibliša we used a bilingual Serbian-English scientific journal, INFOtheca (http://infoteka.bg.ac.rs), covering the field of Library and Information Sciences. A TMX document collection was generated from INFOtheca articles using another of our tools, named ACIDE, an integrated development environment for generating aligned parallel texts [Obradović et al., 2008]. As for available lexical resources, we had at our disposal Serbian morphological e-dictionaries [Krstev, 2008], Serbian and English wordnets (SrpWN and EWN), and a bilingual Serbian-English Dictionary of Library and Information Science technology (further referred to as Dictionary of Librarianship) [Kovačević et al., 2004]. An analysis of results obtained by Bibliša revealed that in some cases the available bilingual resources, namely the wordnets and the Dictionary of Librarianship yielded unsatisfactory results. Hence, a need arose for a third bilingual lexical resource, which we named Biblimir. To that end we developed an additional functionality within Bibliša that allows incremental development of Biblimir.

In the next section we describe the overall system architecture of this tool aimed at using and developing bilingual lexical resources. The third section describes in more detail the components aimed specifically for developing bilingual lexical resources. System modeling was realized using the Unified Modeling Language (UML) and the applications were developed within the Microsoft .Net and MarkLogic environments. The fourth section outlines the usage of the system with examples illustrating the development of Biblimir.

2. SYSTEM ARCHITECTURE
Bibliša is a complex system composed of several modules. Targeted at textual resources in the form of collections of TMX documents and the corresponding metadata, the system uses other

language resources such as grammars in the form of finite automata and transducers, as well as various lexical resources. Bibliša is able to expand search queries both morphologically and semantically, as well as to another language. One type of lexical resources, morphological e-dictionaries, together with the system of rules for compound inflection, finite automata and transducers, represent the basis for morphological expansion of queries. As for semantic and bilingual expansion, the system relies on Serbian and English wordnets and the bilingual dictionary of Librarianship.

The user formulates the initial query as one or more keywords (simple or multiword). If the user so specifies, Bibliša forwards this query for further morphological and semantic expansion, This is essentially handled by a web service (wsQueryExpand.asmx), which is part of the LeXimir software package, a multipurpose tool also developed by the HLT Group [Stanković et al., 2011]. The web service invokes LeXimir's function library LeXimirCore, whose functions expand the query, using available lexical resources and Unitex routines (http://igm.univ-mlv.fr/~unitex). The expanded query is transformed into an XQuery and used for searching the TMX document collection obtained from journal articles. As a result, a set of aligned concordances is obtained, which are presented to the user [Stanković et al., 2012].

3. INTRODUCING BIBLIMIR

As we have already mentioned, the available wordnets and the Dictionary of Librarianship were not sufficiently developed to secure optimal performance of the system. Although it might have seemed that a solution of this problem should be looked for in the refinement of these resources, we opted for the development of a third bilingual lexical resource. The basic reason was that the available resources (except SrpWN) were not developed within the HLT Group and hence had to be used on an "as is" basis. But even the development of SrpWN is restricted to a certain extent by the fact that Serbian synsets have to be linked to corresponding English synsets by the interlingual index (ILI) [Tufiş, 2004]. Another reason is the assessment that a bilingual resource in tune with Bibliša's specifics would substantially contribute to its performance. Hence, Biblimir was born.

3.1 The Logical Model

The logical model of Biblimir, with its characteristic classes, their attributes, as well as relations among the classes is depicted in Figure 1 in the form of an UML class diagram. It draws on basic features of wordnets, but is much simpler in form. The terms in all languages (term list) are modeled by the class *TermEntry*. Based on this list, sets of synonyms (which we will refer to as synsets, as in wordnets) are built, represented by the *TermSynset* class. The relation between these classes is many-to-many, which means that one term can belong to one or more synsets, and that one synset can have several terms from the corresponding language. In its initial phase, Biblimir is conceived as a bilingual Serbian-English resource, but the model enables further expansion to other languages, such as French, German, etc.

Relations between two synsets are modeled by the *SynsetRelation* class, which contains attributes representing foreign keys, as well as the *RelationType* attribute. Currently, this relation is restricted to the 'translational equivalent' relation, which corresponds to ILI in wordnets. However, implementation of other relations is planned, akin to those available in wordnets: hypernym, meronym, antonym, derived, and the like.

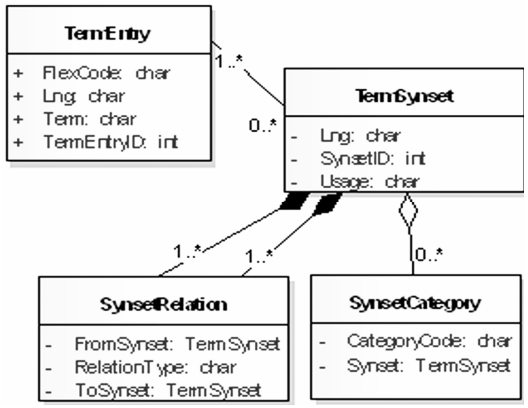

Figure 1: Logical model of Biblimir.

Synsets can be optionally classified into one or more categories such as librarianship or informatics, their subcategories, or categories belonging to other related scientific areas. This classification is modeled by the *SynsetCategory* class using the appropriate category codes. Thus, for example, the synset *impact* is classified into categories inf (informatics) and mng (management), the synset *index* into inf (informatics) and print (printing), whereas the synset *librarian* belong to categories bibl (librarianship) and pers (persons).

3.2 The Object Model

The object UML diagram in Figure 2 offers an insight in the structure of the modeled system through the example of the English term *column*. The diagram should be read from left to right, starting with: *TermEntry, Term=column*. The diagram shows that this term appears in two English synsets, in one of them as the only term {column} and in the other with a second term {column, newspaper column}. Each of these synsets has its corresponding synset in Serbian. Both of them contain three Serbian words. Namely {stub, stubac, kolona} for {column}, and {kolumna, rubrika, novinski stubac}, for {column, newspaper column}. Hence, the diagram displays a total of four synsets, two synset relations and eight terms.

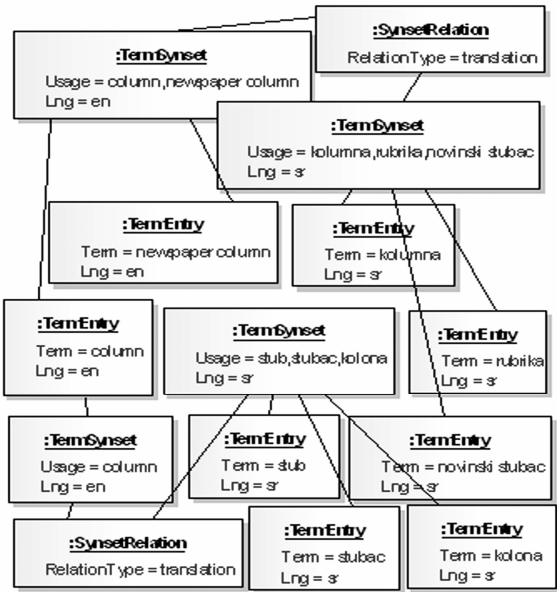

Figure 2: Object model of an example from Biblimir.

3.3 User Interface

Access to Biblimir with functions for its further development is integrated in the Bibliša system, and is available at the following URL: http://hlt.rgf.bg.ac.rs/biblisha/BiblimirSearch.aspx. The user can search this resource by entering a keyword (a term or part of a term) in the search field, selecting a search option: "contains", "starts with" or "exact match", and invoking the search. Matches found in Serbian are presented in a box on the left hand side, while the box on the right shows English terms that result from the search. Usually, one of the boxes is empty, but there are cases when the keyword matches terms in both languages, as for example *printer*. The user can select one or more terms from the two boxes, whereupon a list of corresponding Serbian and English synsets containing the selected term(s) will be displayed. There is also an "Advanced options" button that offers advanced settings, such as selection of categories in which synsets are to be sought.

The "New synset" button allows privileged users to enter new synsets into Biblimir and establish a relation between them. Picture 3 shows the window that opens for this button. Two English synsets {column} and {column, newspaper column} have been entered, together with their Serbian synsets with translational equivalents {stub, stubac, kolona} and {kolumna, rubrika, novinski stubac}. When the user enters a synset, with its terms separated by a comma, the system first checks whether a synset with these terms already exists in its database. If not, the system checks whether all terms from the newly proposed synset already exist in the database and if not, enters the missing terms in the database and generates the new synset. The next step, establishing a relation between a synset and its counterpart in another language, is simply realized by selecting the pair of synsets to be related, and then clicking on the two arrows symbol.

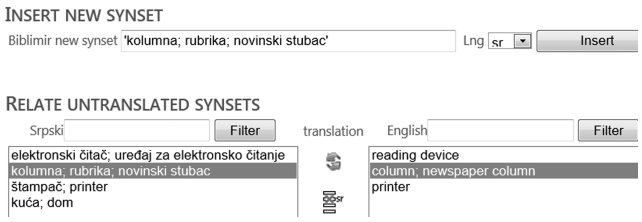

Figure 3: Part of the web page for Biblimir input.

4. DEVELOPING BIBLIMIR

We applied Bibliša to search TMX document collections generated from INFOtheca using available resources. An overview of the size of these resources is given in Table 1. In the course of this search we faced various situations regarding the terms representing translational equivalents in the two wordnets and the Dictionary of Librarianship. Namely, the concordances obtained by the search often revealed the incompleteness of these resources - when the search term appeared in one language without its translational equivalent in the other. In such cases an appropriate entry to Biblimir was considered. In this section we will illustrate this process with several examples.

Table 1: Overview of Available Resources

E-dictionaries	Serbian Wordnet	Dictionary of Librarianship
128,000 simple word lemmas	17,500 synsets	11,300 English terms
10,000 compound word lemmas	30,000 literals	12,100 Serbian terms

4.1 Adequacy of Available Resources

We will start with several examples in which available resources were sufficient and no entry to Biblimir was considered.

Some terms, as for example, the Serbian term *lisni katalog* appear in both SrpWN and the Dictionary of Librarianship. In the latter, the term *lisni katalog* appears together with its synonym *konvencionalni katalog*, whereas *card catalogue, card file, manual catalogue, manually-operated catalogue* appear as English translational equivalents. In SrpWN *lisni katalog* has no synonyms, and its counterpart in EWN is the synset {card catalog, card catalogue}. A search of the collection of TMX documents obtained from INFOtheca articles initiated by the keyword *lisni katalog* and expanded morphologically returned nine concordances that confirmed the adequacy of available Serbian and English terms for this concept. Hence no entry to Biblimir was needed.

When the query was initiated with the English term *public library*, the system did not find its Serbian counterpart in SWN, but it found two corresponding Serbian terms in the Dictionary of Librarianship {javna biblioteka, narodna biblioteka}. The query produced 68 concordances in which the term *public library* was matched with Serbian terms *javna biblioteka* and *narodna biblioteka*. The only exception was in the name of the institution *Narodna biblioteka Srbije*, which was translated as *National Library of Serbia*. Hence, no entry to Biblimir was needed for the term *public library*, except for the compound {Narodna biblioteka Srbije}-{National Library of Serbia}.

4.2 Partial Adequacy of Available Resources

Partial adequacy of available resources was detected mostly in cases when a term appeared in only one resource, or when orthographic variants of the same concept appeared.

English term *user-friendly* appears in EWN, but it does not exist in the Dictionary of Librarianship. On the other hand the Dictionary of Librarianship contains the term *user friendly* (without the dash) and the corresponding Serbian terms are *jednostavan za upotrebu* and *lak za učenje i korišćenje*. The analysis of concordances obtained for the query *user friendly* showed that its Serbian translational counterparts in the Dictionary of Librarianship are not sufficiently precise. Hence, the following list of Serbian terms describing this concept was entered in our new resource, Biblimir: *jednostavan za upotrebu, lak za učenje, lak za korišćenje, prilagođen korisniku, okrenut korisniku*. However, when a new query was initiated with the Serbian term *lak za korišćenje*, the concordances featured the English term *easy to use* as its counterpart, as well as its orthographic variant *easy-to-use.* Hence, the following English synset was entered in Biblimir for the concept in question: {user-friendly, user friendly, easy to use, easy-to-use}.

Similarly, for the Serbian term *ključna reč*, in EWN the corresponding synset is {key word}, whereas in the Dictionary of Librarianship the synset is {keyword}. As both *keyword* and k*ey word* appeared in the 40 concordances obtained for the search initiated with the term *ključna reč*, an entry was made in Biblimir with {ključna reč} as the Serbian synset and {keyword, key word} as the corresponding English synset.

4.3 Incorrect Translational Equivalents

Incorrect translation equivalents appear sometimes in translations of texts, but they can appear in lexical resources as well.

A query initiated by the keyword *browser* revealed that the EWN synset {browser, web browser} has no equivalent in SrpWN, whereas the Dictionary of Librarianship features *pretraživač* as its equivalent. However, *pretraživač* is an incorrect translation for *browser*, the correct one being *prelistač*. Due to this incorrect translation there were no Serbian matches for *browser* in the concordances. Based upon their analysis, the synset pair {browser, web browser}-{prelistač, veb prelistač, pregledač veba} was entered into Biblimir. Query expansion for the keyword *browser* after this addition to Biblimir is depicted in Figure 4.

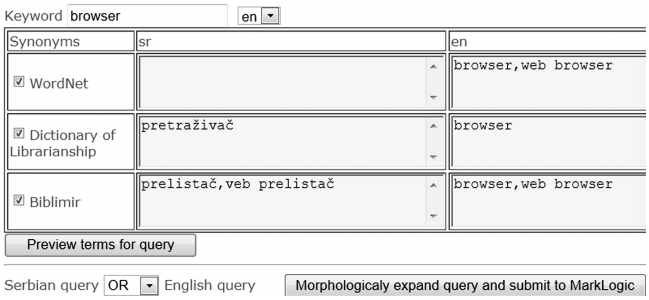

Figure 4: Part of the web page for query expansion.

The English term *repository* appears in both resources: in EWN with the synset {spremište} as its counterpart in SrpWN, and in the Dictionary of Librarianship with the corresponding Serbian synset {depo, skladište}. However, out of 108 concordances obtained for the query initiated by the keyword *repository* only in seven of them the term *repository* was used in the abovementioned meaning. The meaning in which *repository* was used in the majority of concordances corresponds to the Serbian term *repozitorijum*. Hence the pair {repozitorijum}-{repository} was entered into Biblimir. Besides, the concordances also featured English *institutional repository* and Serbian *institucionalni repozitorijum* as translational equivalents, and this pair was also entered into the new dictionary.

4.4 Absence of Terms

In some cases the concordances revealed an absence of adequate terms for a concept in both available lexical resources.

Terms *electronic learning* and *e-learning* and their Serbian translational equivalents *elektronsko učenje* and *e-učenje* do not exist in either of the resources. Hence the English synset {electronic learning, e-learning} and its Serbian counterpart {elektronsko učenje, e-učenje} were entered into Biblimir.

Serbian term *semantički veb* does not exist in available resources. The Dictionary of Librarianship uses English orthography {semantički web}, but in the 15 obtained concordances the term appears only as *semantički veb*. Hence the pair {semantic web}-{semantički veb} was entered into Biblimir.

5. CONCLUSION

Development of a new resource, Biblimir, using a tool aimed at collections of aligned articles in TMX format, Bibliša, has been justified by examples of inadequacy of existing resources, given the fact that these resources themselves could not be enhanced due to proprietary issues. The search performance of Bibliša is expected to improve with the development of Biblimir. The development process itself is semi-automatic, as it needs manual intervention for generating entries for the new resource. In the future, further automatization of this process will be considered, as well as the possibility of adding more languages to Biblimir, thus making it a multilingual lexical resource. As a member of META-NET, the Serbian HLT Group plans to make this resource, among others, available to the larger LT community through META-SHARE.

6. ACKNOWLEDGMENTS

This research was supported by the CESAR (Central and South-East European Resources) project (ICT PSP programme, no. 271022) and by the Serbian Ministry of Education and Science under the grant #III 47003.

7. REFERENCES

[1] Filip, D., Lewis, D., Sasaki, F. 2012. The Multilingual Web: Report on Multilingualweb Initiative. In: Proceedings of the 21st International Conference Companion on World Wide Web (Lyon, France, April 16-20, 2012), pp. 251-254.

[2] Kovačević, Lj., Injac, V., Begenišić, D. 2004. Bibliotekarski terminološki rečnik - englesko-srpski, srpsko-engleski, Beograd: Narodna biblioteka Srbije.

[3] Krstev, C. 2008. Processing of Serbian – Automata, Texts and Electronic Dictionaries. Belgrade: Faculty of Philology, University of Belgrade.

[4] Obradović, I., Stanković, R., Utvić, M. 2008. An Integrated Environment for Development of Parallel Corpora (in Serbian). In: Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen (pp. 563-578), B. Tošović (Ed.). Berlin: LitVerlag.

[5] Pammzi, A., Fabbri, M., Moneglia, M., Zini, M. 2006. Multi-Term Keywords for Indexing Multilingual Textual Repositories: Developing Language Resources and Algorithms. In: Proceedings of the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS '06), (Leeds, UK, December 13-15, 2006), pp. 173-180.

[6] Piperidis, S. 2012. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), (Istanbul, Turkey, May 23-25, 2012).

[7] Stanković, R., Obradović, I., Krstev, C., Vitas, D. 2011. Production of Morphological Dictionaries of Multi-Word Units Using a Multipurpose Tool. In: Proceedings of the Computational Linguistics-Applications Conference, October 17–19, 2011. Jachranka, Poland (pp. 77-84), K. Jassem, P. W. Fuglewicz, M. Piasecki and A. Przepiórkowski (eds.), Polish Information Processing Society, ISBN 978-83-60810-47-7

[8] Stanković, R., Krstev, C., Obradović, I., Trtovac, A., Utvić, M. 2012. A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), (Istanbul, Turkey, May 23-25, 2012).

[9] TMX 1.4b Specification. 2005. http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

[10] Tufiş, D. (ed.) 2004. Romanian Journal on Information Science and Technology. Special Issue on BalkaNet, vol. 7. Romanian Academy, 248 p. ISSN 1453-8245.