RALITSA DUTSOVA

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
r.dutsova@yahoo.com

WEB-BASED DIGITAL LEXICOGRAPHIC BILINGUAL RESOURCES

Abstract

The paper presents briefly a web-based system for creation and management of bilingual resources with Bulgarian as one of the paired language. This is useful and easy to use tool for collection and management of a large amount of different linguistic knowledge. The system uses two sets of natural language data: bilingual dictionary and aligned text corpora.

Keywords: web-application; bilingual resources; lexicographic resources; digital resources; dictionary; aligned corpus; data mining

1. Introduction

The term — “*lexicographic bilingual resources*” usually designates a set of bilingual dictionaries, bilingual corpora which show the usage of words and expressions in two selected languages, and bilingual thesauri which represent conceptual information between two languages. During the last years the language resources are being digitized to make them accessible via internet. The digitization is a long and difficult process, but in this way we receive an opportunity for more effective and efficient preservation, management and presentation of these resources. The development of tools manipulating with lingual resources is a complex and hard process which overcomes difficulties of different nature. The digitization of language resources requires systematization of the data and elimination of the inaccuracies and discrepancies; collecting large language repositories; creation of suitable formal models that adequately reflect the specific features of the natural languages; design and development of a user-friendly interface to access, manipulate, search data and display results in a well systematized and formalized way.

The paper reviews the current state of the web-based software developed to manage bilingual lexicographic resources with Bulgarian as one of the paired language. This web-based software system uses two kinds of linguistic resources: bilingual

dictionary and bilingual corpus. This is a big bilingual collection of special kind of structured texts, which can be used not only for research purposes, but in the daily life for educational and translation purposes.

The web-based system consists of four independent parts (modules): —Dictionary||, —Corpus||, —Search tool|| and —Connection||. Each module is accessible by own interface. The modules —Dictionary|| and —Corpus|| have their own database, own administrator and end-user part. Each database is very well structured and organized according to the text resources that they serve. Special module —Search tool|| was developed, which contains new possibilities for searching and representing the information, stored in the dictionary database. The module —Connection|| links the rest three modules. This link allows interactions between the three independent tools: a parallel search in the dictionary and corpus databases for realization of the cross language information retrieval.

The development of the software system that manipulates utilizes and contains bilingual resources is a complex and long process. In the different period of time the different components have been developed and the modules undergo different changes.

2. Dictionary

The dictionary management system implements the following functions: adding, modifying and deleting entries; adding, changing and deleting elements and attributes, entry search and sorting based on different criteria. The resources used in the dictionary online system can be bilingual with possibility the second language to vary. Currently the ongoing version is using the

Bulgarian-Polish digital dictionary created during the project —Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary||.

In the web-based dictionary system is foreseen functionality representing the homonyms. Homonyms with diverging morphological forms are encoded as separate entries in the lexical database (Dimitrova, Panova, & Dutsova, 2009) and represented in the dictionary as different database records. In the dictionary administrative part on the word-entry page, there is a field where the user must specify a homonym index- a number (Roman numerals) which shows the order of the meaning. The user should insert separate dictionary entry of each meaning of the homonyms. In the end-user part all meanings are listed and ordered according their homonym index.

I зави|я, -еш *vp. event, transitive; zakrećić transitive; skrećić intransitive; ~я се refl. event, intransitive*

II зави|я, -еш *vp. event, intransitive; zawyć intransitive; ку|что ~ от бо|лка*
pies zawył z bólu

3. Corpus

While the bilingual dictionary represents translated words for a word of source language, a bilingual corpus represents a translated sentence for a sentence in the source language (Dimitrova & Dutsova, 2013). Sometimes it is necessary to check the word usage and expression information from the bilingual corpus. The

Въвеждане на глагол						
Индекс за омоним	<input type="text"/>					
Заглавна дума *	<input type="text" value="зави я"/>		<input type="button" value="търси в списък с думи"/>			
2 л. ед.ч. сег. време *	<input type="text" value="-еш"/>	Спрежение на глагола		<input type="text" value="III"/>		
Св. / несп. вид на глагола *	<input type="text" value="св. вид"/>	Преходен / неспреходен глагол		<input type="text" value="преходен"/>		
Събитие / състояние	<input type="text" value="събитие"/>					
добавяне на други грам. категории за думата *						
Значение на полски						
№ група на точни значения *	Значение на полски *	Преходен / Непреходен глагол	Сфера на употреба	Стилистично значение	Латинско значение	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="button" value="добави"/>
1	zakrećić	transitive				<input type="button" value="изтрий"/>
2	skrećić	intransitive				<input type="button" value="изтрий"/>
Деривация/фразеологии/примери на думата						
Вид *	Фраза *	Сфера на употреба	Стилистично значение	Значение на полски *		
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>		<input type="button" value="добави"/>
refl	ня се	event	intransitive			<input type="button" value="изтрий"/>

Figure 1: —Dictionary administrative| panel — adding the first homonym of headword —зави|я| with meaning in English —*to wrap*| (cover or enclose something/something in paper or soft material).

—Corpus| module is independent from the second language. It offers an interface allowing the end-user to search by word in the primary language, selected by the user. All pairs of aligned text where the searched word has been found are listed in a table. The user can search in more than one literary work. This module has its own administrative part. The user can add, edit, delete from and search within the corpus database.

Adding a new aligned text to the corpus database is done through the following steps. If we should add a new Bulgarian-Lang2 aligned text which is Lang2 still doesn't exist in the database, we should add it. If the Lang2 already exists, this step should be skipped. The next step is to add some details concerning the new literary work, which is the primary and the second language, the author's name and the title in both languages. The last step is to upload the text file in the database.

The end-user interface is very simple. It consists of field where the user can enter a free text, word or part of it and drop-down list, where all available in the database literary works are listed.

The result is displayed in a table. All sentences where the search term is met are listed.

4. Search Tool

The implemented dictionary database can be used for extracting, filtering and ordering linguistic knowledge. It is independent from second language and can

Въвеждане на глагол						
Индекс за омоним	<input type="text" value="II"/>					
Заглавна дума *	<input type="text" value="зави я"/>		<input type="button" value="търси в списък с думи"/>			
2 л. ед.ч. сег. време *	<input type="text" value="-еш"/>		Спряжение на глагола		<input type="text" value="III"/>	
Св. / несп. вид на глагола *	<input type="text" value="св. вид"/>		Преходен / непереходен глагол		<input type="text" value="непереходен"/>	
Събитие / състояние	<input type="text" value="събитие"/>					
добавяне на други грам. категории за думата *						
Значение на полски						
№ група на точни значения *	Значение на полски *	Преходен / Непереходен глагол	Сфера на употреба	Стилистично значение	Латинско значение	
<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>	<input type="button" value="добави"/>
1	zawuć	intransitive				<input type="button" value="изтрий"/>
Деривация/фразеологии/примери на думата						
Вид *	Фраза *	Сфера на употреба	Стилистично значение	Значение на полски *		
<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>		
ед	кучето ~ от бо̀лка			pies zawył z bólu		
				<input type="button" value="изтрий"/>		

Figure 2: —Dictionary administrative|| panel— adding the second homonym of headword —зави́я|| with meaning in English —*to yowl*|| (make a loud wailing cry).

речник	корпус	за проекта	поддръжка	вход - регистрация
<div> <div>Български -> Полски</div> <div> <div>зави́я</div> <div>преведи</div> </div> <div> <div>а б в г д е ж з и й к л м н о п</div> <div>р с т у ф х ц ч ш щ ъ ь ю я</div> </div> </div>				
<div> <div> <div>завиван</div> <div>завивка</div> <div>завиден</div> <div>завиждам</div> <div>завинаги</div> <div>зависим</div> </div> <div> <p>I зави́я, -еш <i>св. вид, събитие, преходен, III спряжение; zakrećić transitive; skrećić intransitive; -я</i> <i>св. refl. event, intransitive</i> ;</p> <p>II зави́я, -еш <i>св. вид, събитие, непереходен, III спряжение; zawuć intransitive; кучето ~ от бо̀лка</i> <i>pies zawył z bólu;</i></p> </div> </div>				
<div> <div>Търсене в БГ-ПЛ корпус за:</div> <div> ☐ зави ☐ зави́я ☐ зави́еш </div> </div>				
<div>Произведение: Малкия принц- Антоан дьо Сент-Екзюпери</div>				

Figure 3: —End-user module|| — translation of the Bulgarian verb —зави́я|| to Polish.

be used for any Bulgarian-Lang2 dictionary. For the search tool a new end-user interface has been developed, where different search options have been added. The dictionary casual user interface allows only to search by word. In the search tool we have more options: lemma search- words starting with, ending with or containing

List of uploaded files	
Add new language	
Add new aligned text details	
Upload files	
Add new language	
Language abbrev *	<input type="text" value="BG"/>
Language *	<input type="text" value="PL"/>
<input type="button" value="Save"/>	

Figure 4: —Corpus administrative panel|| — First step: add language.

List of uploaded files	
Add new language	
Add new aligned text details	
Upload files	
Add new aligned text details	
First Language *	<input type="text" value="BG"/>
Second Language *	<input type="text" value="PL"/>
Title aligned text (First language) *	<input type="text" value="Малкия принц"/>
Title aligned text (Second language) *	<input type="text" value="Mały Książę"/>
Author name (First language) *	<input type="text" value="Антоан дьо Сент-Екзюпери"/>
Author name (Second language) *	<input type="text" value="Antoine de Saint-Exupéry"/>
<input type="button" value="Save"/>	

Figure 5: —Corpus administrative panel|| — Second step: add new details for aligned text corpora.

List of uploaded files	
Add new language	
Add new aligned text details	
Upload files	
Add file for aligned corpora	
Select aligned corpora *	<input type="text" value="Малкия принц / Mały Książę"/>
Chapter (only numbers 1-9)	<input type="text"/>
<small>(Leave empty if there are no chapters)</small>	
Please select file to upload: *	<input type="button" value="Browse..."/> <input type="text" value="Le-Petit-Prince-Bg-Pl-aligned.txt"/> <input type="button" value="Upload"/>

Figure 6: —Corpus administrative panel|| — Third step: upload new text file.

concrete string, tag search- to filter the information by part of speech, to search only for derivations, phrases or examples, to extract nouns without singular/plural form etc (Dutsova, 2014). The combination between lemma and tag search is also possible, which we call multiple criteria search. The results are displayed in a table where all words fulfilling the search criteria are listed. The listed words

QUERY IN

скри Search on Bulgarian language in Малкия принц- Антоан дьо Сент-Екзюпери

а б в г д е ж з и й к л м н о п
р с т у ф х ц ч ш щ ъ ь ю я

Result(s): 1

Search proposals: ☒ скри ☐ скрия ☐ скриеш

ID	БГ текст	ПЛ текст
0000000237	Но семената са невидими. Те спят, скрити в земята, докато на някое не му хрумне да се събуди. Тогава се протяга и първо плахо пуска към слънцето едно очарователно безобидно стръкче.	Ale ziarna są niewidoczne. Śpią sobie skrycie w ziemi aż do chwili, kiedy któremuś z nich przyjdzie ochota obudzić się. >Wypuszcza wtedy cudowny, bezbronny pęd, który najpierw nieśmiało wyciąga się ku słońcu.

Search in the dictionary: ☒ скри ☐ скрия ☐ скриеш

Figure 7: End-user query and concordances with Bulgarian headword —**скри**|| (to hide).

are hyperlinks that are references to the corresponding dictionary entries. The retrieved information is very well systematized and structured.

information retrieval tool back to dictionary login

Insert search criteria

C%

Use wildcard character ("%") to substitute a character or characters in a string.
Use semicolon character (";") to enter multiple search criteria.

search Clear form

а б в г д е ж з и й к л м н о п
р с т у ф х ц ч ш щ ъ ь ю я % ;

☒ Verb

Verb conjugation ☐ I ☐ II ☐ III

☒ vi ☐ vp

☒ State ☐ Event

☒ Transitive ☐ Intransitive

☐ Phrases ☐ Examples ☐ Derivations

☐ Noun

Noun gender ☐ f ☐ m ☐ n

☐ Name only in singular

☐ Name only in plural

Nouns with the same plural form

☐ Phrases ☐ Examples ☐ Derivations

☐ Adjective

Female form

Male form

Neutral form

☐ Phrases ☐ Examples ☐ Derivations

Figure 8: User request form: verbs starting with —c||, transitive, expressing state, imperfective aspect.

Figure 9: Search results: list of Bulgarian verbs starting with —c||, transitive, expressing state, imperfective aspect.

Figure 10: Detail view of one of the verbs listed after search has been performed.

The module —Connection|| is just a common end-user interface joining the three independent components —Dictionary||, —Corpus|| and —Search tool|| in order to form a complete system, processing digital bilingual resources with primary language Bulgarian (Dutsova, 2014). The user has the possibility to search for information in the both databases simultaneously and in parallel. Only search by word is available. The results are displayed in a table, where the available information from the both databases is listed. The realization of the mechanism for parallel search in the both databases helps to represent the new functionality in the system for cross language information retrieval. The term —cross-language information retrieval|| is subfield of information retrieval dealing with retrieving information written in a language different from the language of the user’s query. For example, the user’s query is in one language (Lang1) but retrieved relevant

information is in another language (Lang2).

речник | корпус | за проекта | поддръжка | вход - регистрация

Български -> Полски

скрия

☒ Търсене в речник ☒ Търсене в корпус (произведение: Малкия принц- Антоан дьо Сент-Екзюпери)

Търсене

Речник		Корпус							
<p>скрия, -еш св. вид, събитие, преходен, I спрежение; schować, skryć, ukryć, zakryć transitive; zataić transitive; -я се aux. schować się, skryć się, ukryć się, zakryć się</p>		<p>1 резултат: скрия</p> <table border="1"> <thead> <tr> <th>ID</th> <th>БГ текст</th> <th>ПЛ текст</th> </tr> </thead> <tbody> <tr> <td>0000000237</td> <td> <p>Но семената са невидими. Те спят, скрити в земята, докато на някое не му хрумне да се събуди. Тогава се протяга и първо плахо пуска към слънцето едно очарователно безобидно стръкче.</p> </td> <td> <p>Ale ziarna są niewidoczne. Śpią sobie skrycie w ziemi aż do chwili, kiedy któremuś z nich przyjdzie ochota obudzić się. >Wypuszcza wtedy cudowny, bezbronny ped, który najpierw nieśmiało wyciąga się ku słońcu.</p> </td> </tr> </tbody> </table>		ID	БГ текст	ПЛ текст	0000000237	<p>Но семената са невидими. Те спят, скрити в земята, докато на някое не му хрумне да се събуди. Тогава се протяга и първо плахо пуска към слънцето едно очарователно безобидно стръкче.</p>	<p>Ale ziarna są niewidoczne. Śpią sobie skrycie w ziemi aż do chwili, kiedy któremuś z nich przyjdzie ochota obudzić się. >Wypuszcza wtedy cudowny, bezbronny ped, który najpierw nieśmiało wyciąga się ku słońcu.</p>
ID	БГ текст	ПЛ текст							
0000000237	<p>Но семената са невидими. Те спят, скрити в земята, докато на някое не му хрумне да се събуди. Тогава се протяга и първо плахо пуска към слънцето едно очарователно безобидно стръкче.</p>	<p>Ale ziarna są niewidoczne. Śpią sobie skrycie w ziemi aż do chwili, kiedy któremuś z nich przyjdzie ochota obudzić się. >Wypuszcza wtedy cudowny, bezbronny ped, który najpierw nieśmiało wyciąga się ku słońcu.</p>							

Figure 11: Result displayed after the search of Bulgarian word —скрия|| /to hide/ via the component —Connection|| in both repositories of data in Bulgarian — corpus and dictionary.

6. Conclusion

The well-structured lexicographic resources like the digital dictionaries and corpora have a great number of applications: such resources can be integrated in multi-layer databases, can be used for training of translators (for human and machine translation), in linguistics for contrastive studies of the paired languages and for extracting language data, in education for language learning, etc. The realizations of such applications needs a development of specific software tools for processing lexicographic resources in digital form and keep the resources up-to-date. It's not so easy task because the users need effective access to data, advanced and customized tools for search and management the data. Therefore, the described above software system will be useful not only for researchers but also for casual users.

References

- Dimitrova, L. & Dutsova, R. (2012). Implementation of the Bulgarian-Polish online dictionary. *Cognitive Studies / Études cognitives*, 12, 219–229.
- Dimitrova, L. & Dutsova, R. (2013). Web-application for the presentation of bilingual corpora (Focusing on Bulgarian as one of the paired languages). *Cognitive Studies / Études cognitives*, 13, 183–193. <http://dx.doi.org/10.11649/cs.2013.012>
- Dimitrova, L. & Koseska, V. (2009). Bulgarian-Polish Corpus. *Cognitive Studies / Études cognitives*, 9, 133–141.
- Dimitrova, L., Dutsova, R., & Panova, R. (2011). Survey on current state of Bulgarian-Polish online dictionary. In P. Osenova, S. Piperidis, M. Slavcheva, & C. Vertan (Eds.),

- Language Technology for Digital Humanities and Cultural Heritage: Proc. of the International Workshop held in conjunction with the International Conference RANLP — 2011, 16 September 2011, Hissar, Bulgaria* (pp. 43–50). Shoumen: INCOMA.
- Dimitrova, L., Panova, R., & Dutsova, R. (2009). Lexical database of the experimental Bulgarian-Polish online dictionary. In R. Garabík (Ed.), *Metalanguage and encoding scheme design for digital lexicography: Proc. of the MONDILEX Third Open International Workshop, 15–16 April 2009, Bratislava, Slovakia* (pp. 36–47). Brno: Tribun EU.
- Dutsova, R. (2012). Online dictionary — Tool for preservation of language heritage. In P. Stanchev & R. Pavlov (Eds.), *Digital Presentation and Preservation of Cultural and Scientific Heritage: Proc of the International Conference DiPP'2012, 18–21 September 2012, Veliko Tarnovo, Bulgaria* (pp. 142–151).
- Dutsova, R. (2013). Web-application for presentation of Bulgarian language heritage: Bilingual digital corpora and dictionaries. In P. Stanchev & R. Pavlov (Eds.), *Digital Presentation and Preservation of Cultural and Scientific Heritage: Proc. of the International Conference DiPP'2013, 18–21 September 2013, Veliko Tarnovo, Bulgaria* (pp. 99–108).
- Dutsova, R. (2014). Web-based software system for preservation of language cultural heritage. In *Proc. of the International Conference — Digital Presentation and Preservation of Cultural and Scientific Heritage DiPP'2014, Veliko Tarnovo, Bulgaria* (pp. 142–151).
- Dutsova, R. (2014). Web-based software system for processing bilingual digital resources. *Cognitive Studies / Études cognitives*, 14, 45–55. <http://dx.doi.org/10.11649/cs.2014.004>
- Dutsova, R., & Dimitrova, L. (2013). Software system for processing Bulgarian digital resources: parallel corpora and bilingual dictionaries. In K. Gajdošová & A. Žáková (Eds.), *Natural Language Processing, Corpus Linguistics, E-learning: Proc. of 7th International Conference SLOVKO'2013, 13–15 November 2013, Bratislava, Slovakia* (pp. 40–50). Lüdenscheid: RAMVerlag.
- Garabík, R., Dimitrova, L., & Koseska-Toszewa, V. (2011). Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). *Cognitive Studies / Études cognitives*, 11, 227–239.

Acknowledgment

The work, described in this paper, was supported by the PhD Program of the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences.

The author declares that she has no competing interests.

This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 PL License (<http://creativecommons.org/licenses/by/3.0/pl/>), which permits redistribution, commercial and non-commercial, provided that the article is properly cited.

© The Author 2015

Publisher: Institute of Slavic Studies, PAS, University of Silesia & The Slavic Foundation