Web-application for Presentation of Bulgarian Language Heritage: Bilingual Digital Corpora and Dictionaries

Ralitsa Dutsova

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

Abstract. The paper describes three software packages – the main components of a software system for processing and web-presentation of Bulgarian language resources – parallel corpora and bilingual dictionaries. The author briefly presents current versions of the core components "Dictionary" and "Corpus" as well as the recently developed component "Connection" that links both "Dictionary" and "Corpus". The components main functionalities are described as well. Some examples of the usage of the system's web-applications are included.

Keywords: Dictionary Entry, Bilingual Digital Dictionary, Bilingual Digital Corpus, Online Dictionary, Language Resources, Parallel Corpus, Aligned Corpus

1 Introduction

The software packages described in this paper handle great repositories of natural language data in Bulgarian. The main reason for development and implementation of such software tools is Bulgarian language preservation– because the language is a part of the cultural heritage of our country. The tools use two sets of natural language data: bilingual dictionaries and text corpora. The dictionaries are big collections of special structured texts in one or more languages, with a lot of additional information: definitions, etymology, phonetics, pronunciations, examples of usage and other information [1, 4]. A corpus or text corpus is a large set (repository of texts) nowadays usually electronically stored and processed. A corpus may contain texts in a single language (monolingual corpus) or texts in multiple languages (multilingual corpus). Parallel corpus is a set of parallel texts (a text placed alongside its translation or translations). A special kind of corpus is an aligned corpus, it consists of segments, which are original/translated version of the same texts and links/relations between corresponding segments. The result of the alignment of two parallel texts is a merged document, usually called bi-text, composed of both source- and target-language versions of a given text that retains the original sentence order. They have applications in statistical machine translation, contrastive studies and linguistic analysis, for validating linguistic rules in specific researches.

The digitalization of both dictionaries and corpora is costly time-consuming difficult process, which requires creation, editing, cleaning, linking and updating the data. The **software system** used for processing bilingual digital resources is developed at the Institute of Mathematics and Informatics of BAS. The main purpose of its realization is the creation of a useful web tool that allows users to access the available digital language resources at the time. The system actually connects three different software packages each one has its own database and its own user interface.

The first package creates online Bulgarian-Polish dictionary [3], which digital data are accessible by web-application via Internet. The next stage in the process of building the software system is development a web-application for presentation of bilingual aligned corpora with Bulgarian as one of the paired languages [2]. The third component is a package "Connection" that was added on the third phase of development of the software system to realize a connection between both previous components. So a need of a small reprogramming of the both previous tools arises. The "Connection" component is just new interface which is connecting both databases: of the bilingual online dictionary which is already independent from specifying the second language (Bulgarian-Lang2 online dictionary) and bilingual aligned corpora with Bulgarian as one of the paired languages.

The software system for Web-presentations of the Bulgarian language resources is implemented to serve the needs of two main groups of users. Administrators- the people who will manage the system and have special authorization to edit, delete and add new data, and the second type of users are the so-called "end-users" (or casual users) – people who use it, who have limited rights- only to visualize and use the data.

2 "Dictionary" – Web-application for the Presentation of Bilingual Dictionaries (with Bulgarian as the Source Language)

"Dictionary" is a multifunctional software tool for the creation and web-presentation of a bilingual online dictionary with Bulgarian as the source language in the pair.

The dictionary tool is based on a bilingual Lexical Database (LDB). The goal is to design a LDB independent of the second (target) language Lang2. In other words, the source language is Bulgarian and the target language is an unspecified changeable language of translation. The current version uses digital Bulgarian-Polish LDB [6].

2.1 Tasks and Web-services Allocated to the "Administrative (Control)" Panel of the Online Dictionary

- To create a web-based Bulgarian-Lang2 dictionary with capabilities of presenting the dictionary entries as a paper one would, is easy to use and does not require additional "administrator" training, to provide functionality for updating the dictionary content, to provide possibilities to store the information about missing words reported by the "end-users";
- To prepare a User Manual that can be regularly updated.

The software tool offers a user-friendly interface for adding, editing, deleting and searching for words. The access to this panel is restricted and only people who have authorization can access it. After the username and user's password have been entered and verified, the user is redirected to the "administrative (control)" panel. From drop-down list the user chooses what kind of POS will be entered. The web form for entering a dictionary entry was recently reprogrammed in order to be more convenient and useful for the administrators (Fig. 1). The web form shows all specifications necessary for selected POS, the fields for other POS are hidden. When all the information is filled out and the user presses the button "Save", the word is added to the database.

We show an example with a dictionary entry whose headword, the Bulgarian verb "**работя**" /*to work*/ appears in the list of selected Bulgarian headwords. A verb is chosen, because the Bulgarian verbs are the richest POS category with specific characteristics and we can show more functionalities of the web tool. Fig. 1 shows the steps which one "administrator" should follow to upload a new entry to the LDB supporting the Bulgarian-Polish online dictionary.

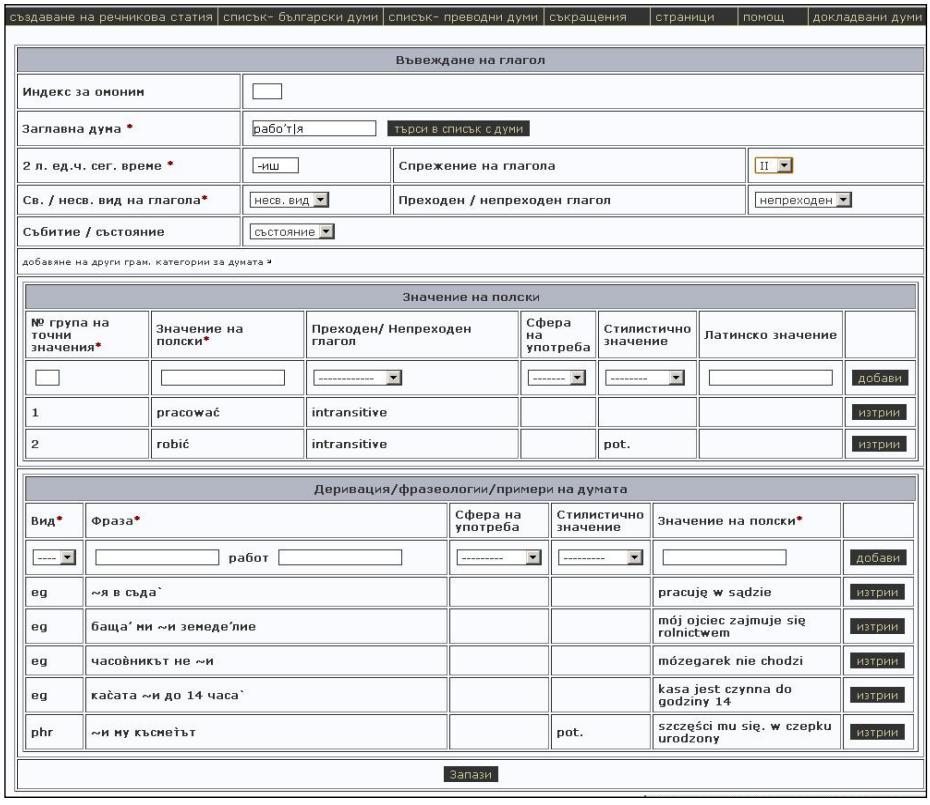

Fig. 1. "Administrative (control)" panel – adding of the grammatical characteristics of the Bulgarian verb "работя" /*to work*/

2.2 Tasks and Web-services Allocated to the "End-user" Part of the Web-site

- To create a user-friendly interface in both languages – Bulgarian and Lang2;
- To ensure quick search of words in the online dictionary;
- To provide accurate and up-to-date information to "end-users";
- To provide the ability for translation in both directions.

The "end-user" part is "bilingual": the user can choose the input language (Bulgarian or Lang2). There is a possibility to search for a translation in both directions: from Bulgarian to Lang2 or from Lang2 to Bulgarian. The translation from Bulgarian to Lang2 will display the whole information existing in the LDB of the dictionary for the searched word. The translation from Lang2 to Bulgarian will be composed only of the main meaning of the Bulgarian headwords [3, 4].

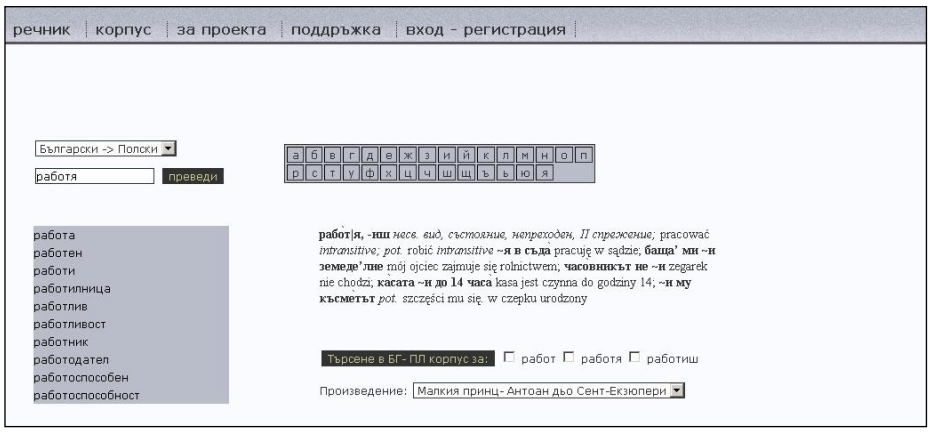

Fig. 2. "End-user" part – displaying the dictionary entry of the verb "работя" /*to work*/ in the end-user part

3 "Corpus" – Web-application for Aligned Corpora Presentation

"Corpus" is a technological tool implemented as a web-based application for the presentation of bilingual aligned corpora with Bulgarian as one the two paired languages. An aligned corpus is a parallel corpus containing relations between corresponding chunks of text of multiple languages [5, 8, 9]. The texts in the ongoing version of the corpora are automatically aligned at the sentence level. We used language-independent freely-available software tools to align bilingual corpora in which Bulgarian was one of the two languages: the MT2007 Memory Translation computer aided tool (TextAlign), and the Bitext Aligner/Converter (Bitext2tmx aligner). Both software packages align bilingual texts without bilingual dictionaries, but human editing is obligatory. The resulting aligned texts (usually called bi-texts) are similar. The main applications of bi-texts are in a computer-assisted human translation, in systems

for machine translation for the training of software tools, for training of translators. The web-based bi-texts can also be used in contrastive studies or other linguistics research, for retrieval of linguistic information, for producing concordances, for developing bi- and multilingual lexical databases and different kinds of digital dictionaries, [7]. The next description focuses to the software tool and user interface.

The component "Corpus" consists of two software packages – an "administrative (control)" panel and an "end-user" part of the web-site. The "administrative (control)" panel has a very simple interface and offers the possibility for the user to add to, edit, delete from and search the database of the corpus.

3.1 Relational Database, Supporting Web-application

The base of the Web-application is the relational database (RDB) of the Bulgarian-Lang2 (Polish in this case) corpora. The relational model is supported by tables containing core information of the corpora entries and the links established between them.

The usage of RDB for storage of the corpora entries has several advantages: maintenance of the integrity of data, ensuring data security and independence, quick and efficient search and data retrieval, upload and updates. We paid special attention to building the database that supports the web presentation of bilingual corpora in order to address the following computational complexities.

Searching a large text can be a costly operation, one that takes up a long time to run. The RDB structure was therefore designed in a way to provide easy and fast search capabilities for the end-users of the bilingual web corpora.

When a user inserts a new record in the RDB through the "administrative (control)" panel, a back-end text parser program takes the input text and simplifies it to its separate constituent words. The different words are then saved in different fields in an index table of the database, and for each word a link is kept to another table where the full text of the aligned pair is saved. This parsing is done for both texts – the one in Lang2 as well as the Bulgarian text. In this way, we achieve a good search performance and only a small delay while inserting new records in the RDB. The delay is not so sensible and the administrator will not pay a big attention to it, because he has the possibility to add the new aligned pairs only one by one. The "administrative (control)" panel provides a simple web-form where the user can insert a new pair of aligned texts [2].

3.2 End-user Web-interface

The web-based end-user interface is bilingual. Only a search-by-word capability is provided to the end user. The user can choose the input language (Bulgarian or Lang2 – Polish, in this case). The search is performed according to the primary language selected by the user. All pairs of aligned text where the searched word has been found are listed in a table. In order to show the word in a better context, together with the target pair we display the previous and next pair as well. If the search results exceed more than 15 records, paging is provided. The Fig. 3 shows the end-user query for

searching Bulgarian word “работя“ /to work/ and resulting concordance with this word in the corpus (aligned texts of R. Kapuściński’s *The Soccer War*).

Fig. 3. End-user query and concordances with Bulgarian word “работя“ /*to work*/

4 “Connection” – Web-application for Dictionary and Corpus Connection

4.1 Main Function of the Connection Component

Administrator & Super Administrator Functions:

- The administrator of any tool has to have access to the “administrative (control)” panel and “end-user” part of the web-site of the other components: that is, access has to be provided from the “administrative (control)” panel of “Dictionary” to the “end-user” part of the web-site and “administrative (control)” panel of “Corpus” and vice versa;
- The administrator has to be able to change personal data for access to the “administrative (control) panel – password and email. If the administrator has access rights

for both “Dictionary” and “Corpus” components, any similar changes have to be noted in both places;
- The super administrator has to be able to create new users or delete existing users. The super administrator has to determine access rights for a given user. From the “administrative (control)” panel of “Dictionary” or “Corpus” it must be possible to create new users (including administrators) who have simultaneous access to “administrative (control)” panels in both components.

User Interface Functions:

- To allow users to search in parallel or in circuit in the dictionary and/or corpus, to display the search results in a synthesised way, to facilitate the user in an effective use;
- The user interface should use not only Bulgarian but also every second language Lang2, loaded in the lexical database;
- Virtual keyboard to facilitate the entry of Bulgarian and Lang2 words;
- A possibility for user registration via a given prompt to provide access to additional functionalities, some of which could be developed in the future.

4.2 Web-application for Corpus and Dictionary Connection

The web-application developed to unite the use of “Dictionary” and “Corpus” tools was easily realized because autonomous user interfaces were developed for each component. Each component is accessible separately, and has its own internet address. The need for developing a common user interface arose with the idea of creating a common system which processes digital bilingual resources with Bulgarian. The “end-user” part of the web-site in the linking modules has a common access, and users will be able to search with it in both databases – dictionary and corpus. A search for Bulgarian and Lang2 words is possible. This component had a relatively simple structure: mirror Bulgarian and Lang2 versions, hyperlinks to the “end-user” part of the web-site of the dictionary and corpus and several sections – “about the project”, “maintenance”, and “entry”. The “Home-page” module consists of a query form with text field, where the user can enter the word of his information search and via a check-box choose where to search. If the user searches for the translation correspondence of the word entered (in the dictionary database), the screen displays the dictionary entry whose headword is this given word. If the user searches the given word in the corpus database, the screen displays the concordance of the given word. A dual search option is also provided that will display on the screen the information present in the dictionary and corpus databases: dictionary entry plus pairs of aligned text where the word occurs.

Since the user interface of “Dictionary” and “Corpus” has a two-way connection for switching between systems, the user is provided with the following possibility: if the query result in any component is “nil”, the user has the possibility to start an analogical search in the other software system by a button click. A small sub-window appears displaying the results of the second search, for example, if the first search was in the dictionary, the sub-window displays the results from the secondary search in

the corpus and vice versa. The new tool will not have its own “administrative (control)” panel. Every component “Dictionary” and “Corpus” has different structures and specifications, so joining them into a single “administrative (control)” panel would create a complex structure accessible via a complex interface and create difficulties for the user.

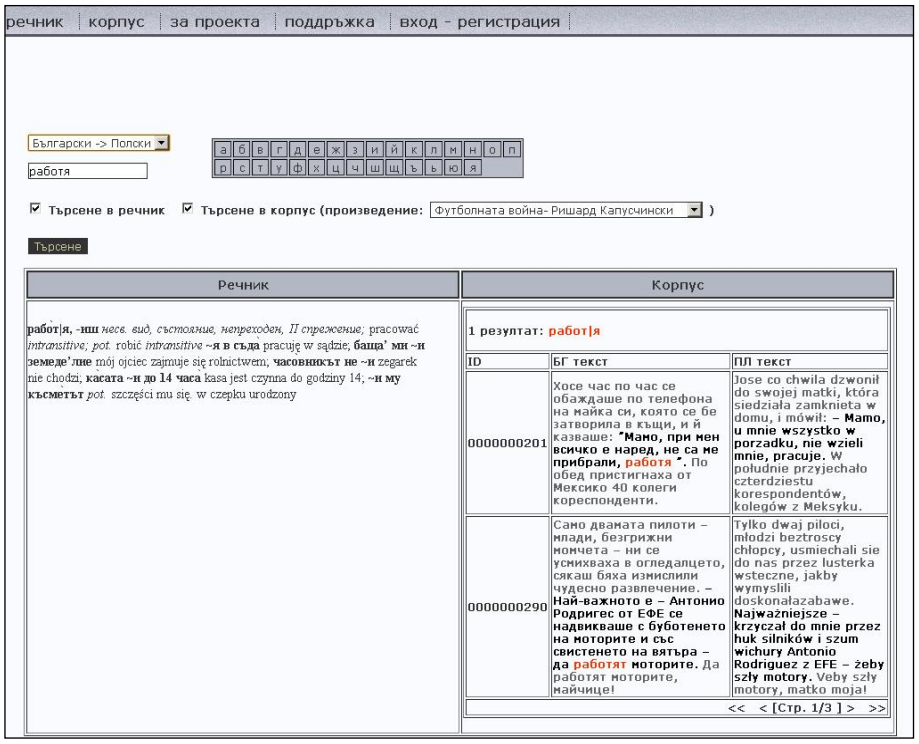

Fig. 4. Result displayed after the search of Bulgarian word “работя” /*to work*/ via the component “Connection” in both repositories of data in Bulgarian – corpus and dictionary.

The only common part between both tools is the login page. When the user loads the system into a web-browser, a login form appears. The login form provides a possibility for the user to enter user-name and password, and then via a radio-button choose which system to enter. After an access rights check, the system redirects the user to the “administrative (control)” panel of the “Dictionary” or the “administrative (control)” panel of the “Corpus”. The user has access to both parts with the same password. However, the “administrative (control)” panels of “Dictionary” and “Corpus” can be used to create users with different access rights: those with access to the dictionary only, and those with access to the corpus only. There may be users with no common simultaneous access to both systems. After the login prompt the system recognizes whether the user is an administrator with full rights and loads only the sections accessible to the user. The “administrative (control)” panel of each component has a link to “administrative (control)” panel and “end-user” part of the web-site of

the other component. If the user wishes to enter the "administrative (control)" panel of the other component, his rights are checked first. If these access rights exist, the user is redirected to the "administrative (control)" panel of the other component and his access rights to the other component are verified. If everything is OK!, a link to the "administrative (control)" panel loads. Thus the user can access the "administrative (control)" panel of the second component without a repeated verification of the access rights. If the user has no access rights to the second component, no link appears. The link to the user interface always loads regardless of the component the user is in and independent of the user access rights.

5 Conclusion

The described software system for processing and web-presentation of Bulgarian language resources (parallel corpora and bilingual dictionaries) is still an experimental tool. The system is intended for research purposes, but it will be applicable in the daily life for educational and translation purposes. The described structures of the tools are still not determined as permanent. Some changes and extensions are possible. Future implementation will include some "search" functions with a query, where the search parameters are fixed and which as a result will extract and show to the user the relevant information.

The main idea of a realization of such system is to enlarge the possibilities of gathering different linguistic knowledge about the natural languages and in particular the Bulgarian language. In order to preserve the natural languages we should have useful and easy to use tools where we can collect and manage the large amount of natural language data.

References

1. Dutsova, R.: Online Dictionary – Tool for Preservation of Language Heritage. In: Proc. of the International Conference "Digital Presentation and Preservation of Cultural and Scientific Heritage", Veliko Tarnovo, Bulgaria, 142-151 (2012)
2. Dimitrova, L., Dutsova, R.: Web-Application for the Presentation of Bilingual Corpora (Fo-cusing on Bulgarian as One of the Paired Languages). J. Cognitive Studies/Études Cognitives. Vol. 13, SOW, Warsaw, 183-193 (2013)
3. Dimitrova, L., Dutsova, R.: Implementation of the Bulgarian-Polish Online Dictionary. J. Cognitive Studies/Études Cognitives. Vol. 12, SOW, Warsaw, 219-229 (2012)
4. Dimitrova, L., Dutsova, R., Panova, R.: Survey on Current State of Bulgarian-Polish Online Dictionary. In: Proc. of the International Workshop "Language Technology for Digital Humanities and Cultural Heritage" within RANLP'2011, 16 September 2011, Hissar, Bulgaria, 43-50 (2011)
5. Dimitrova, L., Garabík, R.: Bulgarian-Slovak Parallel Corpus. In: Proc. of the 6th International Conference "NLP, Multilinguality" SLOVKO'2011, Modra, Slovakia, 20–21 October 2011, 44–50 (2011)

6. Dimitrova, L., Panova, R., Dutsova, R.: Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Proc. of the MONDILEX Third Open International Workshop, 15 – 16 April, 2009, Bratislava, 36-47 (2009)
7. Garabík, R., Dimitrova, L., Koseska–Toszewa, V.: Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). J. Cognitive Studies/Études Cognitives. Vol. 11, SOW, Warsaw, 227–239 (2011)
8. Kelih, E. Slawisches parallel-textkorpus: Projektvorstellung von „kak zakaljalas' stal' (kzs)". In (Kelih, E., Levickij, V., and Altmann, G., eds.), Proc. of the Methods of Text Analysis, 106–124, Černivci. ČNU, (2009)
9. Rosen, A.: In search of the best method for sentence alignment in parallel texts. In (Garabík, R., editor), Proc. of the Third International Seminar "Computer Treatment of Slavic and East European Languages", Bratislava, 10-12 November 2005, 174-185, (2005)