Web-based Software System for Preservation of Language Cultural Heritage

Ralitsa Dutsova

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

Abstract. The paper presents a software system for preservation of Bulgarian language resources – parallel corpora and bilingual dictionaries. The system allows an open access to digital language resources via internet. The main components and the functionalities of the current version of the system are briefly described. The article emphasizes on the description of the module "Search" (a tool for information retrieval and data extraction from a bilingual dictionary).

Keywords: bilingual dictionary, dictionary entry, parallel corpora, information retrieval, data extraction, data mining, language preservation.

1 Introduction

The digital age has had a profound effect on our cultural heritage and the academic research that studies it. Many objects part of the cultural heritage, among them language resources are being digitised to make them accessible to both experts and laypersons. The digitalisation gives an opportunity for more effective and efficient preservation, management and presentation of cultural heritage data. In order to explore and exploit this possibility a need to bring together experts from different fields: cultural heritage study, social sciences and humanities on the one hand, and information technology on the other. Due to a prevalence of textual data in these domains, language technologies have to play a significant role in this challenge. Language technologies help to jump the existing language barriers by offering the potential to analyze texts at advanced levels: to extract information and knowledge not only for the research in humanities and social sciences, but as well as for usage in everyday life for human communication, education (language learning), etc.

The article focuses on a software tool for information retrieval and data mining from a bilingual dictionary, developed as a Web-application, its main components, functionality and user interface. This tool is a part of a system intended to preservation of Bulgarian language heritage. A brief description of whole system is also presented. The described version uses Bulgarian-Polish digital lexical database, supporting Bulgarian-Polish online dictionary, developed in IMI – BAS.

The software system is developed to manage bilingual digital resources with Bulgarian as one of the paired language. The system uses two sets of natural language data: bilingual dictionary and aligned text corpora. Both, the dictionary and the cor-

pus, contain big collections of special kind of structured texts in one or more languages, which will be used in many applications. The web-based system has four independent components (modules). All the components are linked and interactions between them are possible, so they form complex homogeneous system for processing language resources. The independent modules of the system are: "Dictionary" (for creation and management of bilingual dictionaries), "Search tool" (for information retrieval and data mining), "Corpus" (for presentation of aligned corpora). The fourth module is the module "Connection", which links the mentioned three modules as components of independent and autonomous system. The development of this system was a long process, upgraded permanently in the course of time. The first step of the implementation was to develop a specialized database and web-based user friendly interfaces to maintain the Bulgarian-Polish digital dictionary created in the frame of joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary" (between IMI-BAS and ISS-PAS) under the supervision of L. Dimitrova and V.Koseska. Afterwards the modern dictionary writing system was rebuilt to be independent from the second language: so Bulgarian is fixed as a first language in the pair. The information stored in the dictionary database is well-structured and systematized. In order to broaden the usability of the dictionary database a "Search tool" has been developed. It contains new possibilities for searching and representing the information stored in the dictionary database. The dictionary and corpus modules have their own databases and own user interfaces. The module "Connection" links all the components and allows interactions between them, so the search can be performed in both dictionary and corpus databases.

2 Dictionary Module

The main system functions for the creation and management of a bilingual online dictionary include the creation of a modern online dictionary using web-technologies and the provision of possibilities for extending and enrichment of the dictionary entries. So two components — an "administrative" part or a "dictionary management system" and an "end-user" part intended to perform user requests through a user-friendly interface — were developed. The dictionary management system implements the following general functions: adding a new entry, modifying and deleting an entry, alphabetical sorting of the entries. The "end-user" part is bilingual. There are possibilities to search in both directions from Bulgarian to Lang2 or from Lang2 to Bulgarian [5], [6], [7]. The translation from Bulgarian language will display full information, available in the database: all linguistic characteristics of the requested word (headword) such as POS and derivations, all translated meanings with examples and phrases. The translation from Lang2 to Bulgarian will display only information for first translated meaning, available in the database [8].

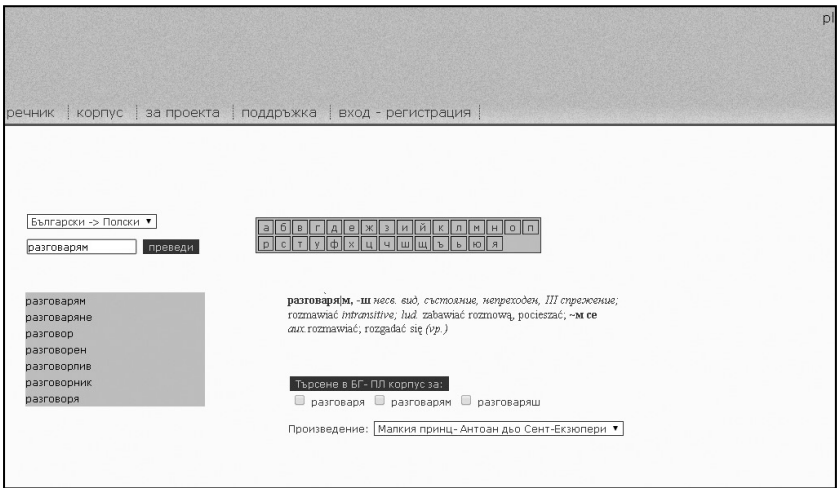

Fig. 1. Dictionary end-user module – translation of the Bulgarian verb "разговарям" /to talk/ to Polish

3 Corpus Module

The module "Corpus" is a technological tool implemented as a web-based application for the presentation of bilingual aligned corpora with Bulgarian as one of the two paired languages [2]. 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 offers the possibility to the user to add, edit, delete from and search within the corpus database. The end-user interface allows search by word in the primary language, selected by the user. The user can search in more than one literary work, currently available in the database, by choosing a title from a drop-down list. All pairs of aligned text where the searched word has been found are listed in a table. The searched word is colored in red in order to emphasize on it. The previous and next pair together with the target pair is displayed as well. At the end of the displayed result table (where the pairs are listed) a link, which redirects the user to the dictionary database, appears - there he can request and perform another search [1], [4].

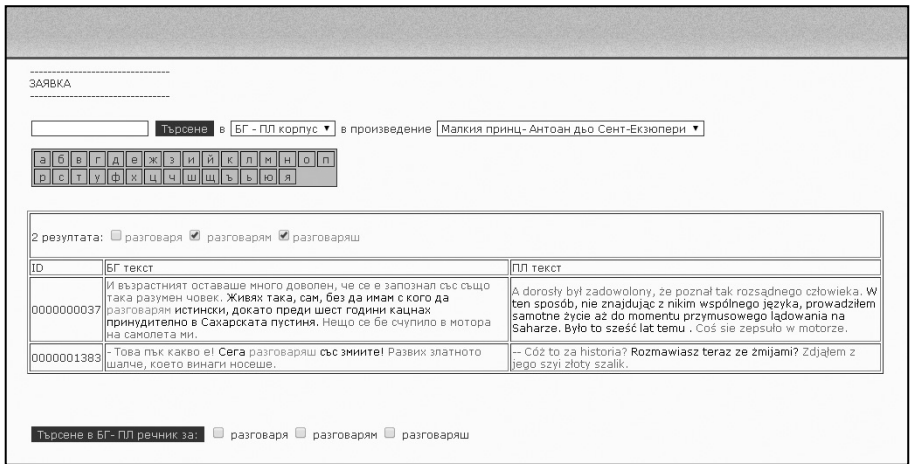

Fig. 2. End-user query and concordances with Bulgarian word " разговарям" /to talk/

4 Search Tool

As we have already to our disposal an implemented relational database for bilingual dictionary [8], the only thing that we need to develop for the "Search tool", is user-friendly interface oriented to the end (i.e. casual) user. This "end-user" interface should provide an effective search in the Web-application database, based on different criteria (given by the users via their requests), to filter the available data; and then to ensure adequate output.

So the main functions of the "Search tool" are:

(1) to process user's requests, i.e. to check the validity of the request and then to search for the requested data,

(2) to produce the results i.e. to extract requested data and to show them to the user's screen.

The end-user module is generally accessible to the casual users. But the user can register by filling in the registration form. The tool enables registered users to save different search criteria and filters (most preferable or usable), so that the user can use them without entering them again. Multiple criteria search is allowed in the "Search tool". The user can search for all words available in the database starting with, ending with or containing concrete string (we call this lemma search), to filter the information by part of speech verbs, nouns and adjectives (we call this tag search), to search only for derivations, phrases or examples, etc. The combination between lemma and tag search is also possible.

Rhymer procedure: If the user enters an initial syllable or a final syllable of a given lemma (so called "rhyme"), a "Rhymer procedure" will produce result as a dictionary of "rhymes". In this case the "Rhymer procedure" retrieves information for the rhymes of a corresponding word. We recognize two types of rhymes: head-rhyme and end-rhymes. Words with head-rhyme have the same initial syllable. Words with end-rhyme have the same final syllable. For example, if the user enters the word

вятър “wind” under this option, “Rhymer” retrieves a list of words ending the same way (e.g. *пъстър* “motley”, *театър* “theatre”, *филтър* “filter”, *хитър* “sly”, etc.). This option lets easily find exact rhymes.

When the casual user loads the Web-application to work with, a web form is loaded: the user can specify there the search type. In order to check the validity of the user requests some control functions in the search procedure are added. In the text field the user can insert lemma, or part of lemma, or a list of several lemmas separated with semicolon. The displayed results can be narrowed by choosing the additional criteria in the web form of the request. The user can specify his/her requirements concerning the words (the lemmas listed in the text field) by clicking selected menu buttons of the web form.

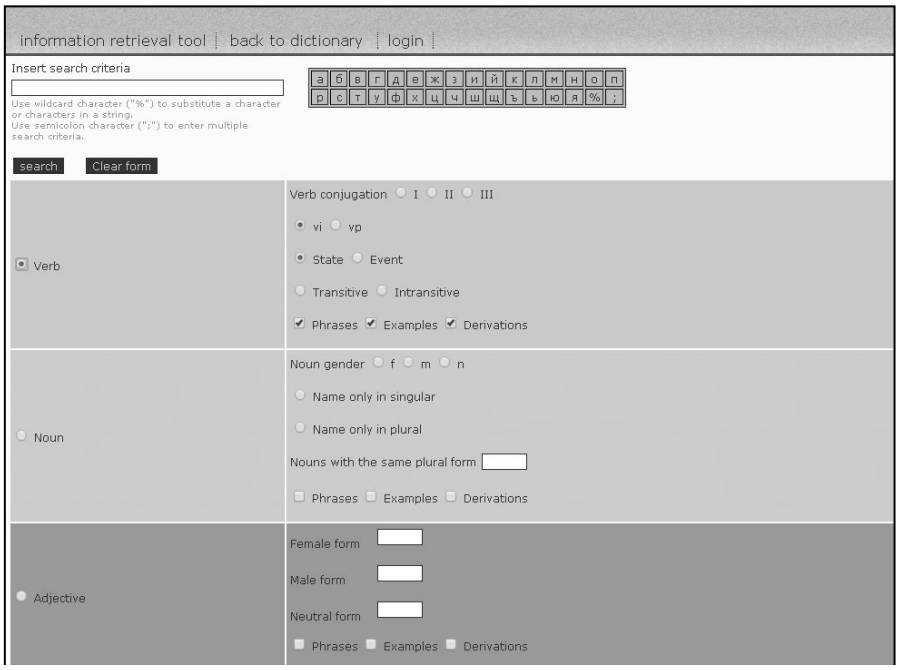

Fig. 3. User request form for searching for and extracting of words /verbs, imperfective aspect, expressing state/

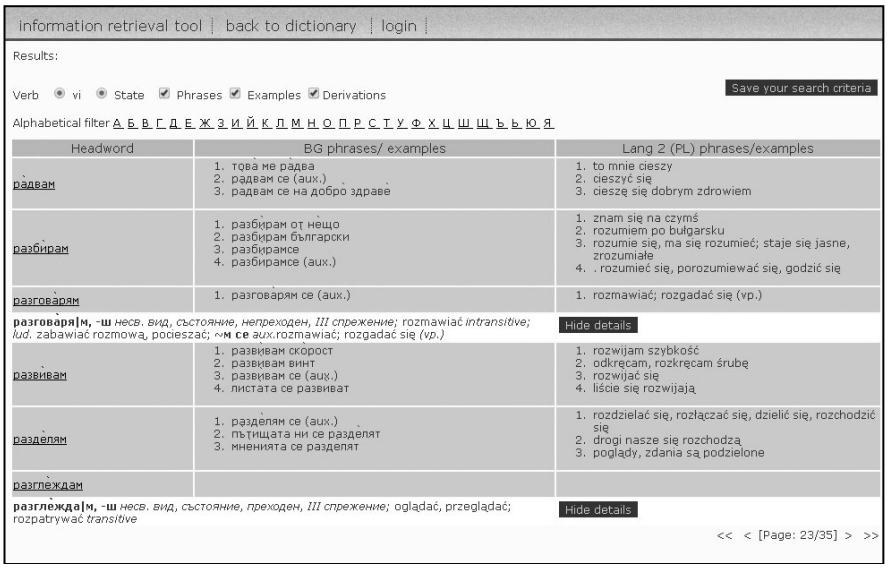

Fig. 4. Bulgarian transitive verbs, imperfective aspect, expressing state

5 Connection

The module “Connection” has been easily developed. Its main goal is to join the dictionary and corpus functionalities, so the user can search in both modules simultaneously. The need to develop a common user interface arose with the idea to create a homogeneous system which processes digital bilingual resources with Bulgarian. The user has the possibility to “see” the information from the both databases, which is very well structured and systematized [1], [4]. The “Home-page” module consists of a query form with a text field where the user can enter the word of his information search and choose where to search via a check-box. The “Connection” 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.

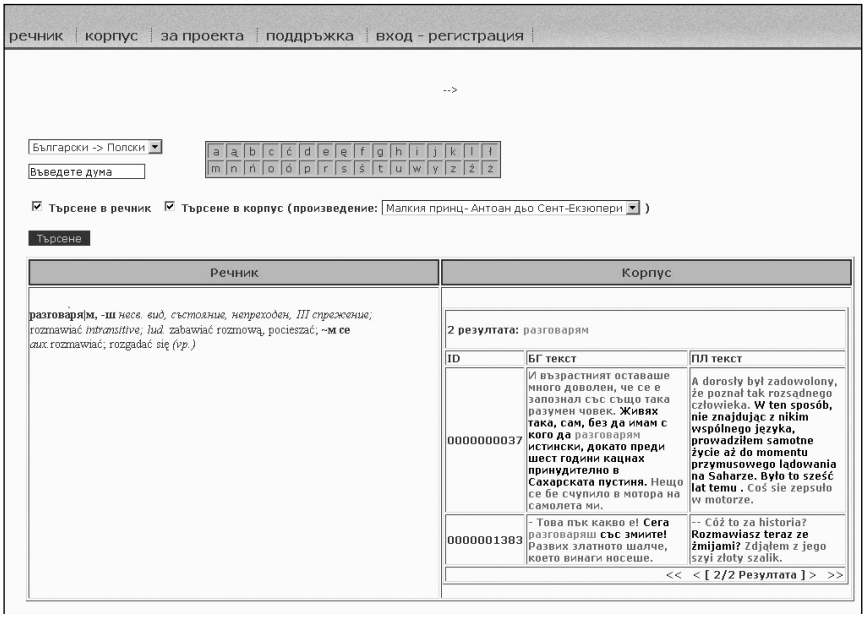

Fig. 5. Result displayed after the search of Bulgarian word “разговарям” /to talk/ via the component “Connection” in both repositories of data in Bulgarian — corpus and dictionary

6 Conclusion

The paper presented briefly a system for creation and management of bilingual resources with Bulgarian. The main idea of the implementation of such system is to enlarge the possibilities of gathering different linguistic knowledge about the natural languages and in particular the Bulgarian lan-guage. 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. (2014), Web-based Software System for Processing Bilingual Digital Resources. In J. Cognitive Studies/Études Cognitives. Vol. 14, SOW,pp. 45-55, Warsaw, Poland
2. Dimitrova, L., R. Dutsova (2013), Web-Application for the Presentation of Bilingual Corpora (Focusing on Bulgarian as One of the Paired Languages). In J. Cognitive Studies/Études Cognitives. Vol. 13, SOW, pp. 183-193 , Warsaw, Poland
3. Dutsova, R., L. Dimitrova (2013), Software System for Processing Bulgarian Digital Resources: Parallel Corpora and Billinagual Dictionaries. In: Proc. of the Seventh International Conference SLOVKO’2013 Natural Language Processing, Corpus Linguistics, E-learning, 13-15 November 2013, pp. 40-50 ,Bratislava, Slovakia

4. Dutsova, R. (2013), Web- application for Presentation of Bulgarian Language Heritage: Bilingual Digital Corpora and Dictionaries. In: Proc. of the International Conference "Digital Presentation and Preservation of Cultural and Scientific Heritage, pp. 99-108 , Veliko Tarnovo, Bulgaria
5. Dutsova, R. (2012), Online Dictionary – Tool for Preservation of Language Heritage. In: Proc. of the International Conference "Digital Presentation and Preservation of Cultural and Scientific Heritage, pp. 142-151 , Veliko Tarnovo, Bulgaria
6. Dimitrova, L., Dutsova, R. (2012), Implementation of the Bulgarian-Polish Online Dictionary. J. Cognitive Studies/Études Cognitives. Vol. 12, SOW, Warsaw, 219-229
7. Dimitrova, L., R. Dutsova, R. Panova (2011), 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, pp. 43-50 , Hissar, Bulgaria
8. Dimitrova, L., R. Panova, R. Dutsova (2009), Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Proc. of the MONDILEX Third Open International Workshop, 15 – 16 April, 2009, pp. 36-47 , Bratislava, Slovakia