Web-based System for Digital Presentation, Management and Preservation of Bulgarian Language Heritage

Ralitsa Dutsova

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

Abstract. The paper briefly describes a web-based software system for presentation, processing and management of Bulgarian language resources as a part of the Bulgarian cultural heritage. The system will be available in the cyberspace. It allows an open access through the global network to the well-structured digital language data - bilingual dictionary and parallel corpora. The structure and main functionalities of the system, implemented as a set of web-applications, are presented.

Keywords: bilingual dictionaries, parallel corpora, language resources, information retrieval, data extraction, data mining, intangible cultural heritage, Bulgarian language.

1 Introduction

The natural languages, oral traditions and expressions, performing arts, rituals and festival events, knowledge and practices concerning nature and the Universe are considered as intangible cultural heritage. Preservation (also "safeguarding") means to ensure the viability of the intangible cultural heritage, including the identification, documentation, research, protection, transmission, particularly through formal and non-formal education, as well as the revitalization of the various aspects of such heritage. The advent of digitization gives new trends and possibilities so we speak mostly about digital preservation (safeguarding) of the intangible cultural heritage. Digitization gives more efficient preservation, management and presentation of the cultural artifacts. The language resources - developed and saved as big repositories - are very often digitalized in order to be made easy accessible via the global network.

The natural languages are part of the intangible cultural heritage. Their digitization is hard, time-consuming and long process. It needs to bring together different experts: from one hand, experts from social sciences and humanities and information technology specialists, from the other. This process can be divided in to two steps. The first step consists in the preparation of the language information, and includes the collection of different kind of language resources, their digitization and updating as digital data and their presentation in a suitable formal model in order to be machine readable. On this step the help and intervention of linguists is required. The second step is the process of development of different software tools, providing easy management and

open access to the language and linguistic information. Specialists as software architects and developers are needed it on this stage.

The web-based software system for safeguarding the language in-formation, described in the article, will serve as a specialized platform to maintain bilingual digital resources with Bulgarian as one of the paired language. The system is focused on two sets of natural language data: bilingual dictionary and aligned text corpora. Both, dictionary and corpora contain complex information. The presentation of this language information in a good formal model ([3], [9]) consists in a long process, requiring intervention of linguists. The implemented system has four completely independent components (modules, developed as web-applications) on one hand, but on the other hand the interaction between them is foreseen [1]. The web-applications are: "Dictionary", "Corpus", "Information Retrieval Tool" and "Connection".

This article elaborates two issues of digital preservation of language heritage: preservation and management of the language and linguistic information of digital dictionaries and corpora; connection between the dictionary and the corpus in order to retrieve semantic information and implementation of information retrieval.

2 Dictionary module

Online bilingual dictionary with Bulgarian as a source language and inde-pendent second language, using web-technologies, is created [6], [7]. The part of language resources, used in the module "Dictionary", namely Bulgarian-Polish resources, was created in the frame of the 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).

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 the two paired languages [3]. Text corpora provide large databases of naturally-occurring discourse, enabling empirical analyses of the actual patterns of use in a language. The strengths of the corpora are illustrated with respect to three areas of research: (1) natural languages grammars; (2) lexicography; and (3) language usage for specific purposes; register variation of words usage. The third research area is designed to tackle the problems faced by a variety of first- and second-language users (specialized translators, undergraduates, junior and experienced researchers, and language trainers).

4 Information Retrieval Tool

The web-search tool uses the database of the implemented already dictionary with Bulgarian fixed as source language, and Lang2 as target one. This tool provides a new way of searching in the dictionary database depending on the user request. The need-

ed information is displayed in a well-systematized list [1], [2]. The new functionality attached to the database of bilingual dictionary allows data mining and extraction of the linguistic information stored in it. Depending of the user search criteria the search tool will retrieve structured linguistic information. Such examples are: “display only the nouns with male/female/neuter gender”, or “display only transitive verbs expressing state”, or “extract verbs in imperfective aspect, expressing state”, “extract words starting with /ending with or containing any string”. All kind of combinations of user search criteria are possible and the user can extract different kind of linguistic information.

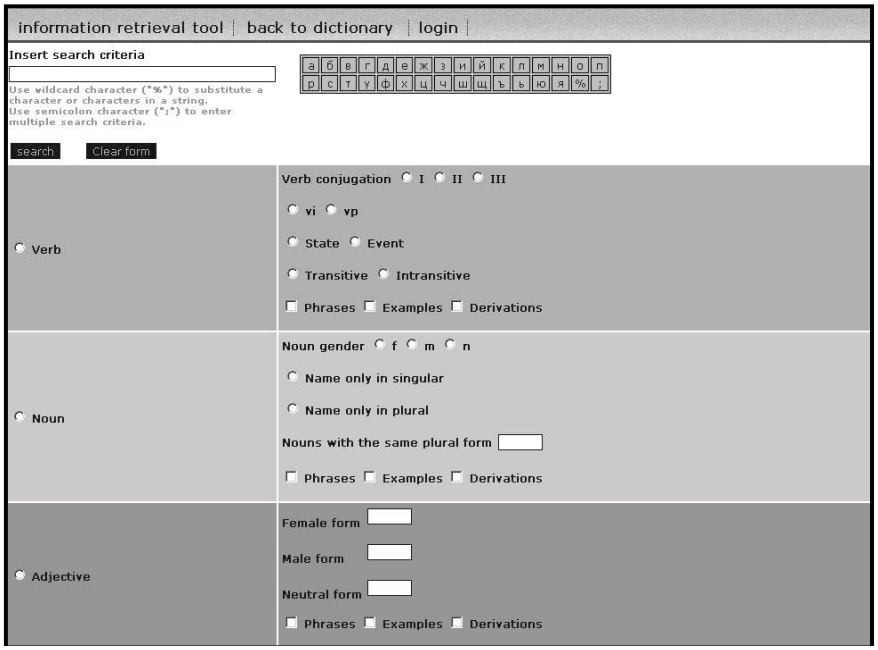

Fig. 1. User request form of the Information Retrieval Tool

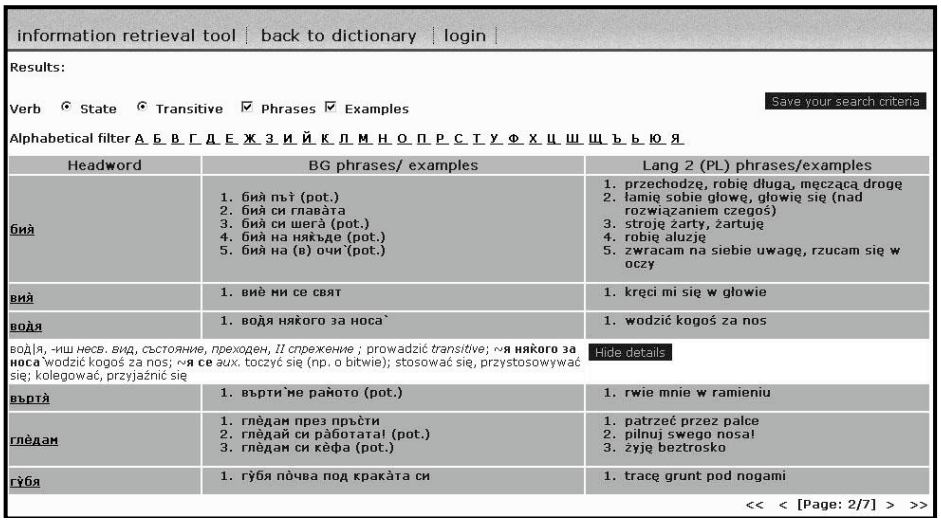

Fig. 2. Result of search of Bulgarian transitive verbs, conjugation II, expressing state, appearing in phrases and examples of dictionary entries

5 Connection

The last step in creating the software system for processing, maintaining and preserving the language resources is the development of the module "Connection". Common end-user interface for joint use of the "Dictionary" and "Corpus" modules is implemented [1], [2]. The idea to develop an additional component arose with the need to search in the dictionary and corpus databases simultaneously and retrieve the language knowledge contained in the both databases. The "Home-page" of the "Connection" application consists of a query form where the users can set their search criteria. The results are listed. The screen presents information extracted from the both databases: a dictionary entry and pairs of aligned texts where the word occurs.

Links switching between the modules "Dictionary" and "Corpus" are fore-seen. If the query result in any component is "NULL", the user could start a new search. A small sub-window will appear, displaying the results of the second search, for example, if the first search was in the dictionary, the sub-window will display the results from the secondary search in the corpus and vice versa.

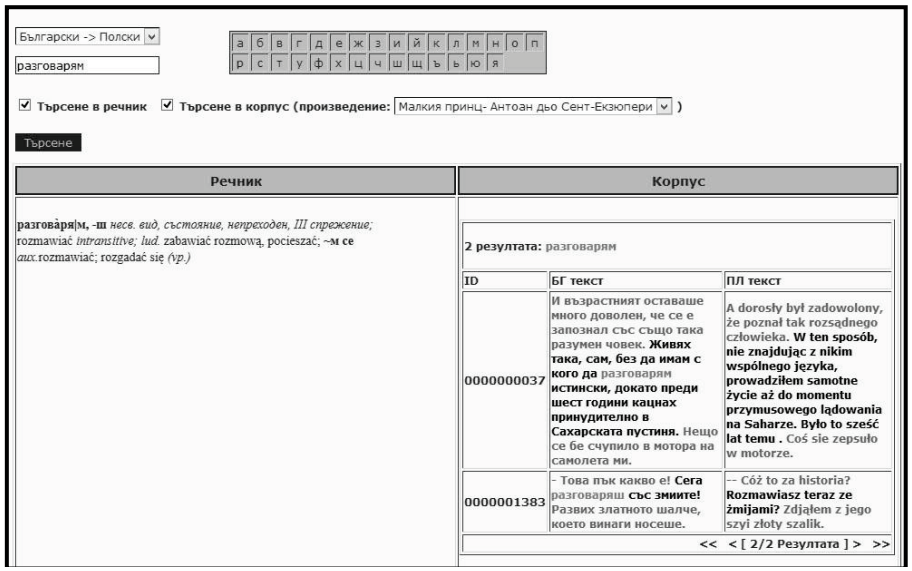

Fig. 3. Result displayed after search of Bulgarian word “разговарям” /to talk/ in both modules “Corpus” and “Dictionary”

6 Conclusion

The software management system for safeguarding the language heritage implements the following general functions: adding (compiling) a new entry; modifying an existing entry: adding/changing/deleting elements and attributes; deleting an entry; entry search based on various features: element/attribute existence; alphabetical sorting of entries. Each component is independent and used separately from the others. The “Dictionary” and the “Corpus” module have their own administrative part in order to be managed independently. Different users can have different rights for access to the complex system.

The web based system for preservation and management of linguistic heritage is perforce in order to collect, to manage, to preserve, to manipulate different kind of language knowledge. The system will be very useful and valuable for translators (human and machine), high school and university students, as well as for every-day users.

References

1. R. Dutsova, (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”, pp. 165-172 , Veliko Tarnovo, Bulgaria

2. R. Dutsova, (2014): Web-based Software System for Processing Bilingual Digital Resources. In: J. Cognitive Studies/Études Cognitives. Vol. 14, pp. 45-55, SOW, Warsaw, Poland
3. L. Dimitrova, 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
4. R. Dutsova, D. Dimitrova, (2013). Software System for Processing Bulgarian Digital Resources: Parallel Corpora and Bilingual Dictionaries. In: Proc. of the Seventh International Conference Natural Language Processing, Corpus Linguistics, E-learning *SLOVKO'2013*, 13-15 November 2013, pp. 40-50 , Bratislava, Slovakia
5. R. Dutsova, (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
6. R. Dutsova, (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
7. Dimitrova, L., Dutsova, R. (2012): Implementation of the Bulgarian-Polish Online Dictionary. In: J. Cognitive Studies/Études Cognitives. Vol. 12, SOW, Warsaw, pp. 219-229
8. L. Dimitrova, 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, Hissar, Bulgaria, pp. 43-50
9. L. Dimitrova, 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, Bratislava, Slovakia, 2009, pp. 36-47