Digital Presentation of Bulgarian Language Heritage - Tools and Web-applications

Ralitsa Dutsova

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

Abstract. The article describes a model of integrating language techniques with leading-edge technology to develop a web-based on-line tool for heritage language preservation and e-learning. Bulgarian language resources - parallel corpora and bilingual dictionary are presented in an interactive way using a common web-interface which is accessible for students, faculty, and affiliated Heritage language communities.

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

1 Introduction

Digitization of language resources provides efficient means for their preservation, management and presentation. Language carries valuable information about the society and culture of its speakers. If the language disappears, important cultural knowledge will also disappear. During the past years a lot of effort has been focused on research and development of a different kind of language preservation tools and software systems. The digitization process is hard and time-consuming. First, language and linguistic information should be well investigated and analyzed so it can be presented in a suitable machine readable formal model. Then, a good software solution should be found in order to assure the best way of presentation, management and preservation of the resources.

The software program presented in this article could be used as a tool for language preservation and e-learning. The software system will ensure access to resources of materials, and will provide heritage learners more opportunities to learn and investigate further the language. This is a special bilingual collection of structured texts (bilingual dictionary and bilingual corpus with Bulgarian as one of the paired language), which can be used not only for research purposes, but in the daily life for educational and translation purposes. The web-based system represented here has four different components: “Dictionary”, “Corpus”, “Search tool” and “Connection”. They have been developed over different time periods and still undergo changes in order to improve their functionalities.

The current article describes briefly the four independent components and illustrates how they form a homogeneous system that manages language resources. It will focus on the technologies used for development of the web system as well.

2 Description of the Web-application

The development of the web-based system that manipulates, utilizes and contains bilingual resources is a challenging process for both linguistic and IT specialists. First, the language resources should be well selected, so they represent in a best and correct way the linguistic knowledge. Then they should be structured in a good formal model in order to be machine readable. The language resources should be used and displayed in different ways without any impact on them. The web-based application will be used as a repository of a large collection of language resources. The linguistic information should be easily displayed and maintained with the help of an interactive and useful interface, which requires different IT specialists and technologies.

This software system allows open access to digital language resources via the Internet. It uses two sets of natural language data: a bilingual dictionary and aligned text corpora. Each of the implemented modules is accessible by its own interface. The "Dictionary" and "Corpus" module have their own database, own administrator and end-user part. The nature of the different text resources requires each database to be structured and organized respectively in a different way. The "Connection" module allows a parallel search in the dictionary and corpus databases and implementation of cross-language information retrieval. The "Search tool" provides just a new way of searching in the dictionary database depending on the user request.

From another point of view the web-application can be considered as a system that consists of two – an administrative and an end user - layers. The administrative layer is used to maintain the databases, to keep the language information up-to-date, easy to correct the data or to add new records. The "Search tool" and "Connection" module do not need to have their own administrative part. They only consist of end user interface.

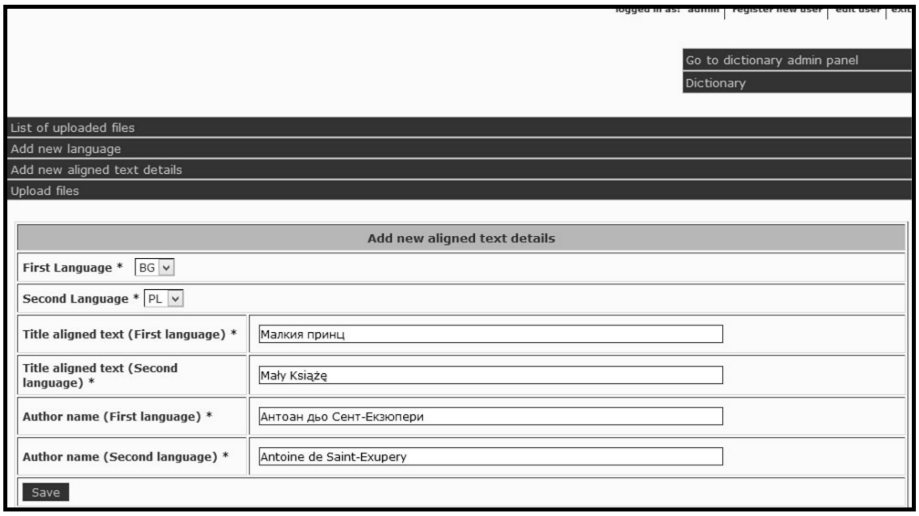
The following screenshots display some of the functionalities of the administrative panel of the "Corpus" module. With the help of the administrative module the user can upload different literary texts directly from a file. All files are kept in the database. When the end user requests some kind of a query, the system checks all uploaded files for the word in the query.

The development of the web-based system that manipulates, utilizes and contains bilingual resources is a challenging process for both linguistic and IT specialists. First, the language resources should be well selected, so they represent in a best and correct way the linguistic knowledge. Then they should be structured in a good formal model in order to be machine readable. The language resources should be used and displayed in different ways without any impact on them. The web-based application will be used as a repository of a large collection of language resources. The linguistic information should be easily displayed and maintained with the help of an interactive and useful interface, which requires different IT specialists and technologies.

This software system allows open access to digital language resources via the Internet. It uses two sets of natural language data: a bilingual dictionary and aligned text corpora. Each of the implemented modules is accessible by its own interface. The "Dictionary" and "Corpus" module have their own database, own administrator and end-user part. The nature of the different text resources requires each database to be structured and organized respectively in a different way. The "Connection" module allows a parallel search in the dictionary and corpus databases and implementation of cross-language information retrieval. The "Search tool" provides just a new way of searching in the dictionary database depending on the user request.

From another point of view the web-application can be considered as a system that consists of two – an administrative and an end user - layers. The administrative layer is used to maintain the databases, to keep the language information up-to-date, easy to correct the data or to add new records. The "Search tool" and "Connection" module do not need to have their own administrative part. They only consist of end user interface.

The following screenshots display some of the functionalities of the administrative panel of the "Corpus" module. With the help of the administrative module the user can upload different literary texts directly from a file. All files are kept in the database. When the end user requests some kind of a query, the system checks all uploaded files for the word in the query.


logged in as: admin | register new user | edit user | exit
Go to dictionary admin panel
Dictionary
List of uploaded files
Add new language
Add new aligned text details
Upload files
Save

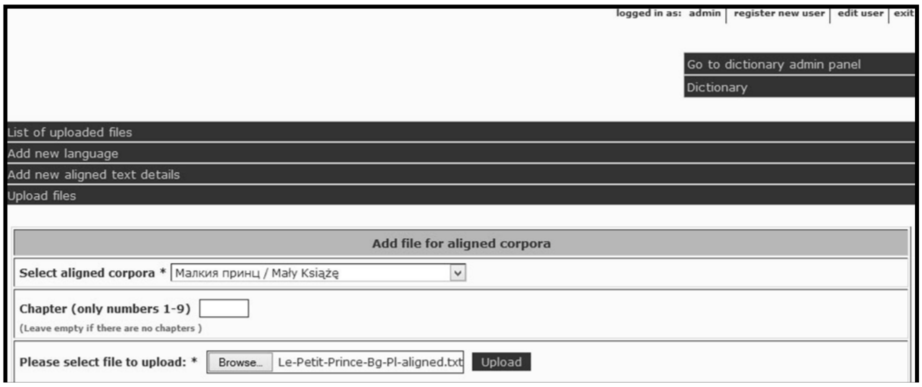
Add new aligned text details

First Language * BG

Second Language * PL

Title aligned text (First language) *	Малкия принц
Title aligned text (Second language) *	Mały Książę
Author name (First language) *	Антоан дьо Сент-Екзюпери
Author name (Second language) *	Antoine de Saint-Exupery

Fig. 1. "Corpus" module administrative layer – adding information for the literary texts uploaded in the system


logged in as: admin | register new user | edit user | exit
Go to dictionary admin panel
Dictionary
List of uploaded files
Add new language
Add new aligned text details
Upload files
Browse...
Upload

Add file for aligned corpora

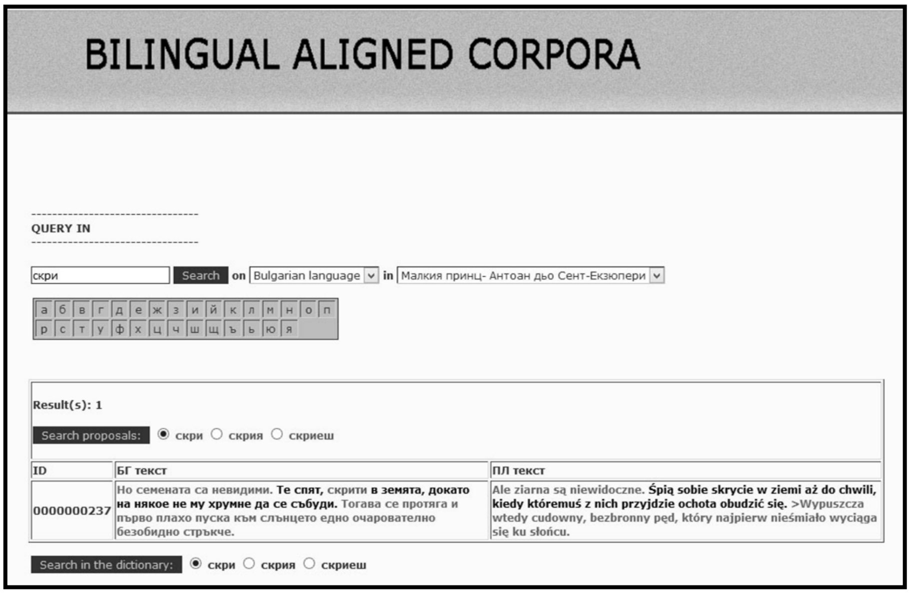
Select aligned corpora * Малкия принц / Mały Książę

Chapter (only numbers 1-9)
(Leave empty if there are no chapters)

Please select file to upload: * Le-Petit-Prince-Bg-Pl-aligned.txt

Fig. 2. "Corpus" module administrative layer – upload of new file

The administrator uploads the literary texts. The necessary information is saved in the database. The collection of aligned corpora in the web system increases. The end users can request different search queries in the correspondent literary text in "Corpus" or "Connection" modules.


BILINGUAL ALIGNED CORPORA
Search
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
Search proposals:
Search in the dictionary:

QUERY IN

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

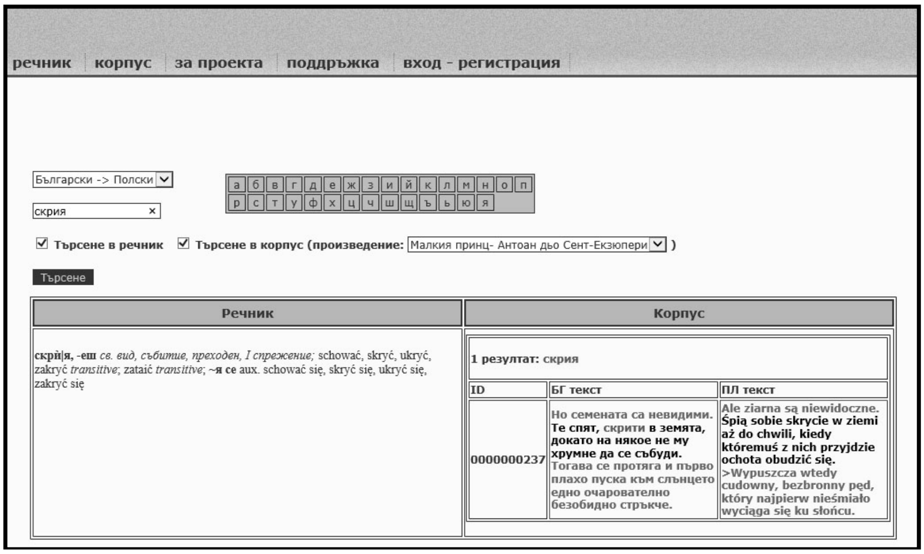
Result(s): 1

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

Fig. 3. "Corpus" module end user layer – a result from the search query performed in the uploaded literary texts



Български -> Полски

скрия

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

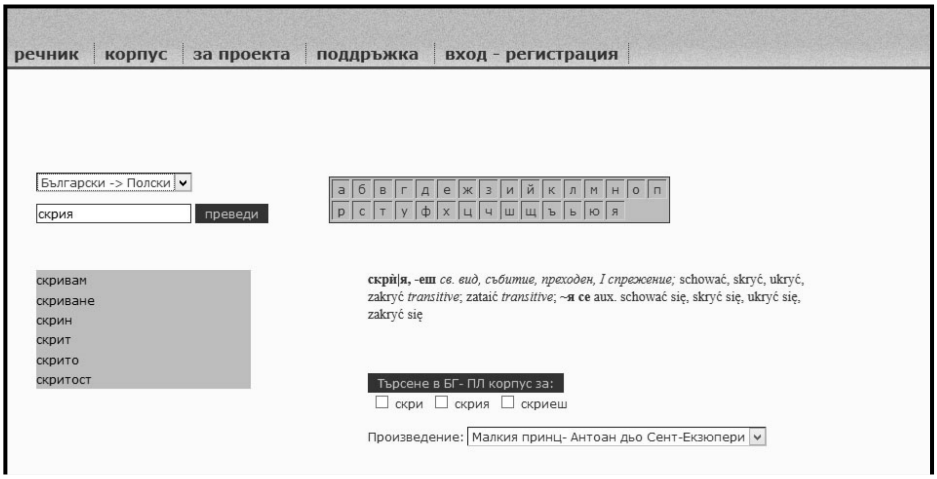
скрѝ|я, -еш *св. вид, събитие, преходен, I спрежение;* schować, skryć, ukryć, zakryć *transitive*; zataić *transitive*; **~я се** aux. schować się, skryć się, ukryć się, zakryć się

1 резултат: скрия

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.

Fig. 4. "Connection" module end user layer – a result from the search query performed in the uploaded literary texts and dictionary database

Directly from the "Connection" module or "Corpus" the user can be redirected to a query in the "Dictionary" or "Search tool" module.


речник
корпус
за проекта
поддръжка
вход - регистрация
преведи
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
Търсене в БГ- ПЛ корпус за:

Български -> Полски

скрия

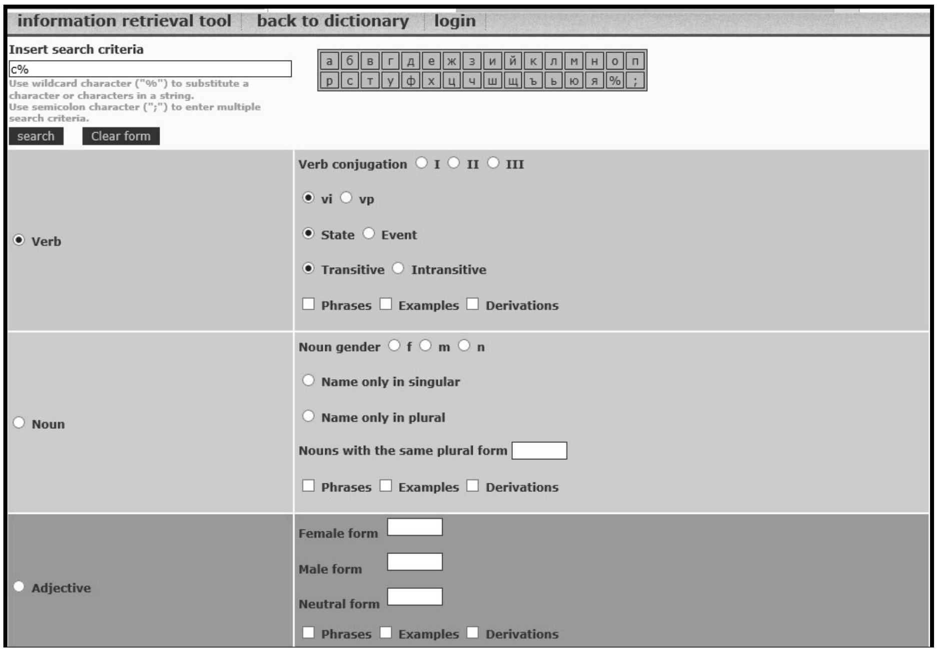
скривам
скриване
скрин
скрит
скрито
скритост

скрѝ|я, -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

☐ скри ☐ скрия ☐ скриеш

Произведение: Малкия принц- Антоан дьо Сент-Екзюпери

Fig. 5. "Dictionary" module end user layer – a result from the search query in the dictionary database


information retrieval tool
back to dictionary
login
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 % ;
search
Clear form

Insert search criteria

c%

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

● 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

Fig. 6. "Search tool" end user layer – additional search criteria for the dictionary database

3 Technologies Used to Implement the Web-application

The web-application's implementation uses free technologies originally designed for producing dynamic web pages with multiple functionalities – Apache, MySQL and PHP.

Apache/MySQL/PHP is a solution stack that is most commonly used for creating and managing web-applications. Apache is a web server, MySQL is a database, and PHP is a scripting language. The three technologies are extremely popular. The most significant reason for their widespread use is that all components are open source. Before uploading the software application in the web, we need to develop and test it locally on desktop computers. It is sometimes hard to set and parameterize them to work together, which is why a lot of software packages were developed to install and set up Apache, MySQL, PHP on a Windows operation system and allow to run web applications locally. This is very useful for development and hence made these three technologies preferable among any other.

The Apache HTTP Server, colloquially called Apache, is the world's most used web server software. Apache is developed and maintained by an open community of developers. It is a secure, efficient and extensible server that provides HTTP services.

Many web-based programs and content management systems need to store and retrieve data. For example, the dictionary entries and the aligned corpus text must be stored in databases. Instead of reinventing and implementing its own system of storing and retrieving data, this web-application uses the database program MySQL.

MySQL is a popular choice of database for use in web applications, and supports large databases containing around 50 million records and several different character sets as well. For example, the Latin diacritics and the Cyrillic letters are not lost when they are entered or retrieved from the database. All data is saved in the chosen character set. Multilanguage support is available. The performance issue is also very important while developing web applications. Most MySQL servers have query caching enabled, one of the most effective methods of improving performance. When the same query is executed multiple times, the result is fetched from the cache, which is quite fast.

PHP is a server-side scripting language designed for web development but also used as a general-purpose programming language. PHP code may be embedded into HTML code, or it can be used in combination with various web template systems, web content management systems and web frameworks. PHP is a programming language with a wide functionality: it is able to evaluate form data sent from a browser, build custom web content to serve the browser, communicate with the database. One of the strongest and most significant features in PHP is its support for a wide range of databases. Writing a database-enabled web page is incredibly simple using one of the database specific extensions e.g. MYSQL. With PHP, we have the freedom of choosing an operating system and a web server. Furthermore, we have the choice of using procedural programming or object oriented programming (OOP), or a mixture of both.

4 Conclusion

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

References

1. R. Dutsova, (2015) Web-based System for Digital Presentation, Management and Preservation of Bulgarian Language Heritage. . In: Proc. of the International Conference “Digital Presentation and Preservation of Cultural and Scientific Heritage, 28 – 30 September 2015, pp. 189-194 , Veliko Tarnovo, Bulgaria
2. R. Dutsova, (2014) Web-based Software System for Processing Bilingual Digital Resources. In: J. Cognitive Studies | Études Cognitives. Vol. 14, SOW, pp. 45-55, 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 SLOVKO’2013 Natural Language Processing, Corpus Linguistics, E-learning, 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 DiPP’2012”, 18 - 21 September 2013, 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, 18 - 21 September 2012, 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, 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, pp. 43-50, Hissar, Bulgaria
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, 2009, pp. 36-47, Bratislava, Slovakia