Online Dictionary – Tool for Preservation of Language Heritage

Ralitsa Dutsova

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

Abstract. The paper aims to represent a bilingual online dictionary as a useful tool helping preservation of the natural languages. The author focuses on the approach that was taken to develop compatible bilingual lexical database for the Bulgarian-Polish online dictionary. A formal model for the dictionary encoding is developed in accordance with the complex structures of the dictionary entries. These structures vary depending on the grammatical characteristics of Bulgarian headwords. The Web-application for presentation of the bilingual dictionary is also described.

Keywords: Annotation Tag, Dictionary Entry, Digital Dictionary, Information Technologies, Lexical Database, Online Dictionary, Relational Database

1 Introduction

The term ‘cultural heritage’ refers to any item relating to the culture or history of a group of people. It includes monuments and historical sites, artifacts, manuscripts, and even abstract phenomena with cultural significance, such as folk tales or the language used by a society. The acknowledgement of the natural language as a part of the cultural heritage provokes for development of robust and flexible solutions for its preservation and presentation. One of the ways to preserve the cultural heritage is to digitize the collections and make them accessible via the Net. The dictionaries can be considered as great repositories of data. They consists a big collection of words in one or more languages with a lot of additional information: definitions, etymologies, phonetics, pronunciations, examples of usage and other information. The digitalization of the dictionaries requires cleaning, linking and enriching the data which is very long and difficult process and leads to develop new concepts and models for management of such kind of specific data.

The Web-based dictionaries propose many advantages. One of them is that they are easy for use and accessible from everywhere if the application is installed on web server and it has its own URL address.

Development of the digitized dictionaries is a complex and hard process which need to overcome many difficulties. One of them is linked to the lack of enough formal models allowing the words to be classified to different language classes and also tools for automatic assignment of the words to the correspondent class.

The main difficulty in realization of design model for the bilingual dictionaries is when the translation of the word should work in the both directions. The lexical forms have more than one meaning and they are not overlapping during the translations in both directions. This makes the realization of the data base supporting the digitized dictionaries logically complex and formidable.

Further in the article we are going to describe the model of the Lexical database used for the realization of the Bulgarian-Polish online dictionary and its web implementation. In the examples we will focus on the verbs as they are the richest from linguistic point of view lexical category in the Bulgarian language.

2 Bilingual Lexical Database Supporting Bulgarian-Polish Online Dictionary

A Lexical database (LDB) is the base of the dictionary implementation. The standards for text annotation SGML and XML are used for design and creation of the Bulgarian-Polish LDB. The lexical information is represented in a hierarchical tree structure. The defined XML- tags which are the nodes of the tree structure contain information for the spelling, phonology, syntax and the semantic of the words.

2.1 CONCEDE Model for Dictionary Encoding

The project CONCEDE has built lexical databases in a general-purpose document-interchange format, for six Central and East European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian. The CONCEDE LDB model offers a standardized hierarchical tree-structure of a dictionary entry with an understandable semantics [6]. The main goal of the project is to produce common guidelines and principles for encoding dictionaries so they are compatible with other TEI conformant resources, to provide a Document Type Definition (DTD) with as few elements as possible each with an unambiguous, clearly defined interpretation [7]. To achieve the goal the task has been divided in two parts: to create both structural and content annotation tags. The structural tags have very well defined heritage semantics and they are the notes from the hierarchical tree-structure of the dictionary entry. The structural tags are tree – **entry** (indicates dictionary entry), **struc** (indicates separate independent hierarchical structure: division of meanings and sub-meanings, examples and phrases), and **alt** (alternation: one word can have more than one spelling). The content tags give information about spelling, phonetic, syntax and the semantic of the words. Main content tags are **hw** (headword, the ***head word*** of the dictionary entry, which will be used for indexing and alphabetical sorting), **pos** (indicates corresponding part of speech class of the head word: verb, noun, adjective, pronoun, etc.), **def** (definition(s) of the head word sense(s)), **eg** (example with the corresponding sense of the head word).

The first LDB for Bulgarian language [5] has been developed under the CONCEDE project and it is based on the Bulgarian Explanatory Dictionary [1].

The example below illustrates how the Bulgarian verb "**живея**" /*live*/ is encoded in the CONCEDE LDB:

```
<entry>
<hw>жив|ея</hw>
<gram>нсв.</gram><subc>нпрх.</subc>
<struc type="Sense" n="1">
<def>Съществувам като организъм: дишам, храня се,
усещам.</def>
<eg><q>Да живееш сто години!</q></eg>
<eg><q>Колко години живее орелът?</q></eg>
<eg><q>Къде живеят рибите?</q></eg>
</struc>
<struc type="Sense" n="2"><usg
type="register">прен.</usg>
<def>Оставям спомен, витая, пребъдвам.</def>
<eg><q>Ботев живее в сърцата на цял народ.</q></eg>
<eg><q>Да живееш сто години!</q></eg>
</struc>
<struc type="Sense" n="3">
<def>Прекарвам времето си някъде или по някакъв
начин.</def>
<eg><q>Живял съм и в село, и в град.</q></eg>
<eg><q>Живея богато.</q></eg>
</struc>
<struc type="Sense" n="4">
<def>Поминувам, прехранвам се.</def>
<eg><q>Живеем от труда си.</q></eg>
<eg><q>Живея с една заплата.</q></eg>
</struc>
<struc type="Phrases">
<struc type="Phrase" n="1"><orth>Да живее!</orth>
<def>Да живее! – да пребъде, да се
слави.</def></struc></struc>
<struc type="Phrase" n="2"><orth>Живеем си.</orth>
<def>Сговаряме се, в добри отношения
сме.</def></struc></struc>
<struc type="Phrase" n="3"><orth>Живея си живота.</orth>
<def>Живея добре, наслаждавам се на
живота.</def></struc></struc>
</entry>
```

2.2 Bulgarian-Polish LDB

The LDB of the Bulgarian-Polish online dictionary follows the CONCEDE model for dictionaries encoding but with some modifications and extensions. First, the monolingual model of the CONCEDE LDB was extended to a bilingual one. Second, some new tags were added to the LDB to cover the specific features of Bulgarian and Polish words aiming more adequate presentation of both Slavic languages. In the realization of the online dictionary we use the same convention as in the CONCEDE LDB for the content and structural tags. The structural tags and some of the content tags in the Bulgarian-Polish LDB are the same as they are defined in CONCEDE model: **entry**, **struc**, **alt** and **hw**, **pos**, **def** and **eg**.

There are no tags in the CONCEDE model for presentation of bilingual data. So, to introduce the Polish translation(s) of the Bulgarian headword, the LDB was extended with new content tag **trans**.

Traditional printed dictionaries did't describes all the specific characteristics of the Bulgarian verbs. To enrich the data in the online dictionary we set an objective to try to represent all the characteristics of the Bulgarian verbs as completely as possible [4, 3]. For this reason several new content tags were added: tag **conjugation** and attribute **type** to represent the conjugation of verb and its type (I, II or III); tag **semantic** and attribute **type** to represent semantic information (type can contain values "1" for verbs expressing "**state**" or "2" for verbs expressing "**event**"). The tag **gram** gives information about *perfect* or *progressive aspect* of verb, and tag **subc** – about its *transitivity* or *intransitivity*.

Example: The entry with headword Bulgarian verb "**живея"** /*live*/ in printed Bulgarian-Polish dictionary [8] is presented as follows;

жив|е'я, -еш *vi. state, intransitive* żyć *intransitive*; mieszkać *intransitive*; **да ~е'е!** niech żyje!; **аз ~е'я в Со'фия** mieszkam w Sofii

The entry with head word "**живея"** in the Bulgarian-Polish LDB is more representative:

```
<entry>
<hw> жив|е'я </hw>
<conjugation>
  <orth>-еш</orth>
  <type>1</type>
</conjugation>
<semantic>
  <orth>състояние</orth>
  <type>1</type>
</semantic>
  <subc> непреходен </subc>
  <pos>v</pos>
<gram>несвършен вид</gram>
```

```
<struc type="Sense" n="1">
  <trans>żyć</trans>
  <subc> intransitive </subc>
</struc>
<struc type="Sense" n="2">
  <trans> mieszkać </trans>
  <subc> intransitive </subc>
</struc>
<struc type='Phrases'>
  <q>да ~е'е!</q>
  <trans>niech żyje!</trans>
</struc>
<eg>
  <q>аз ~е'я в Со'фия</q>
  <trans>mieszkam w Sofii</trans>
</eg>
</entry>
```

3 Relational Database, Supporting Bulgarian-Polish Online Dictionary

The base of the Web-application is the relational database (RDB) of the Bulgarian-Polish dictionary. The RDB is designed with the help of the lexical database (LBD). The relational model is supported by tables containing core information of the dictionary entries and the links, established between them. The direct transformation of the LBD into the RBD ensures all the advantages, provided by the modern relational databases management systems (DBMS). The usage of RBD for storage of the dictionary entries has several advantages: maintenance of integrity of data, ensuring data security and independence, quickly and efficiently search and data retrieval, data upload and update.

The RDB provides also a direct conversion of the dictionary entry into relational format, without passing through the LDB. The possibility of translation from Polish to Bulgarian is also provided, but only Bulgarian headword as a corresponding translation will be visualized: if a casual (end-) user tries to search for the word **mieszkać** (one of Polish translations of the Bulgarian head word "**живея**" /*live*/), the online dictionary will display only the Bulgarian meaning and the information stored in the **subc** tag (*transitive* or *intransitive verb*). For the above example, the screen will show

mieszkać intransitive живея

An XML parser is created to transform the lexical database into the relational database. The aim of the syntactic analyzer is to initialize the relational database that serves as a basis of the dictionary. Then the saved in the RDB entries can be edited through the administrative module of the online dictionary (point 4.1 below).

4 Web-based Application for Representation of the Bulgarian-Polish Online Dictionary

The main goal of the implementation of the Bulgarian-Polish online dictionary is to create an up-to-date dictionary using the web technologies and to foreseen the possibilities to extend and enrich the dictionary entries. The technologies used for the implementation of the web- based application are Apache, MySQL, PHP, AJAX and JavaScript. We use free technologies originally designed for developing dynamic web pages with a lot of functionalities.

The software package for Web-presentation of the Bulgarian-Polish dictionary is implemented with the HTML (Hypertext Markup Language) and CSS (Cascading Style Sheets), two of the core technologies for building Web pages. The package consists of two main modules: administrative module and end-user module.

4.1 Administrative Module of the Bulgarian-Polish Online Dictionary

The administrative module is used to manage the data: insert new entry, delete or update already an existing one. This module is intended only for authorized persons: administrators – users who have the right to introduce new words, to delete or change words existing already in the database, as well as to update and modify the contents of the web-pages in the end-user module. Usually these are people with linguistic knowledge, without any programming skills, trained to work with the administrative module. Such decision ensures that the data in the dictionaries are correct and up-to-date.

How the administrative module works? To explain its function, let's focus our attention on the section "Create dictionary entry" (this section contains more functionalities than all the rest sections of the Web-application).

After the authorized user has been logged in, he is redirect to a page where he can choose from drop-down list what kind of word will be inserted in the database of the dictionary: noun, adjective, verb or other parts of speech (which include adverb, interjection, union, communion, particle, pronoun, preposition and postposition). Let's choose the verb "**живея**" /*live*/. In the MS WORD-format Bulgarian-Polish dictionary this verb is presented as follows:

жив|е'я, -еш *vi. state, intransitive* żyć *intransitive*; mieszkać *intransitive*; **да ~е'е!** niech żyje!; **аз ~е'я в Со'фия** mieszkam w Sofii

Our first step here is to choose "**глагол**" /*a verb*/:

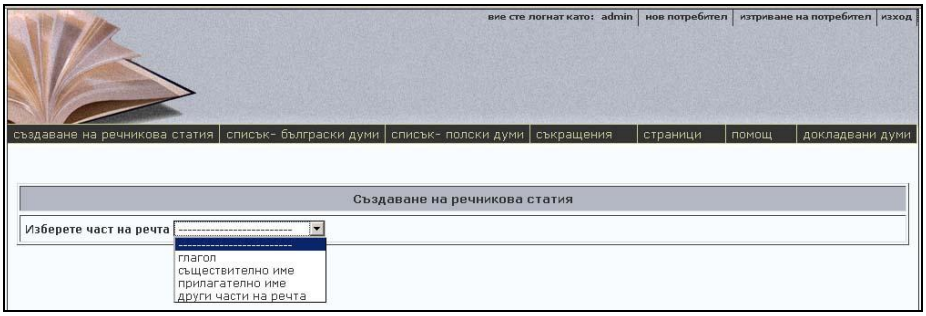

Fig. 1. Administrative module – choosing the type of the new word to be added

The second step of entering a verb is to fill in the headword and its grammatical characteristics (Fig. 2). The conjugation type I, II or III , the transitivity/intransitivity of the verb, the 'perfect aspect' (vp) or 'imperfect aspect'(vi) of the verb and the semantic information 'state' or 'event' can be chosen from the drop-down lists:

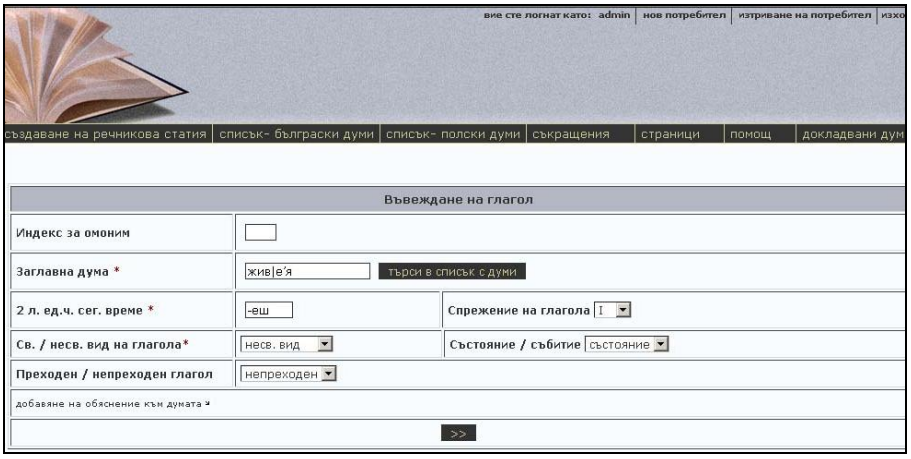

Fig. 2. Administrative module – adding the grammatical characteristics of the verb "**живея**"

When all the information is filled we press the button Next **">>**". On the next step we shell fill in a form with additional information (if it is needed for the specific use of the verb) such as medical term, botanical term, etc. and/or any stylistic meanings (archaic, folklore, etc.). On this step the user can create references to another word. For our example we don't have what to fill in this step so we just skip it and go the next step.

On the third step there is possibility to fill in the Polish meanings of the corresponding Bulgarian headword. With the button "add" we can add many Polish meanings. A drop-down lists which can give detailed information for usage of the Polish verbs is also added. For the Polish verbs we can obtain classifier "transitive or intransitive" as well.

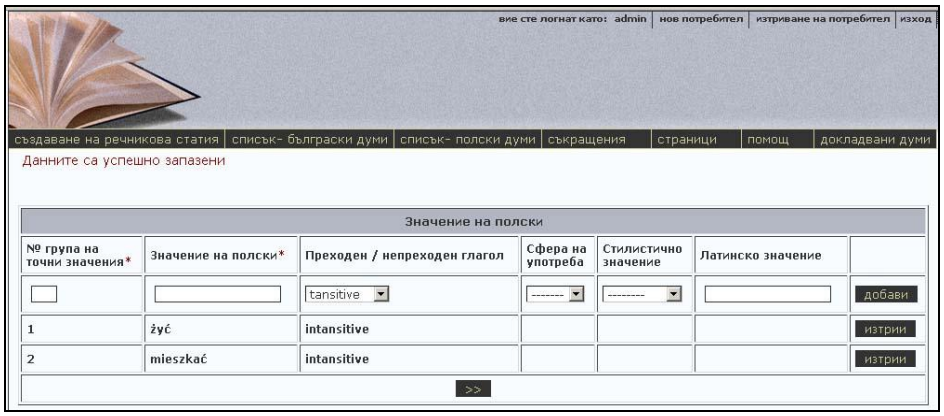

Fig. 3. Administrative module – adding the Polish translations (meanings)

There is common part for each part of speech that ensures the possibility to add unspecified number of derivations, phrases and examples for each headword. This is the last step of entering the verb. When we press on the button "Край" ("Finish") the word stores in the data base. Then it will be possible to search this word and display it by the user-end module.

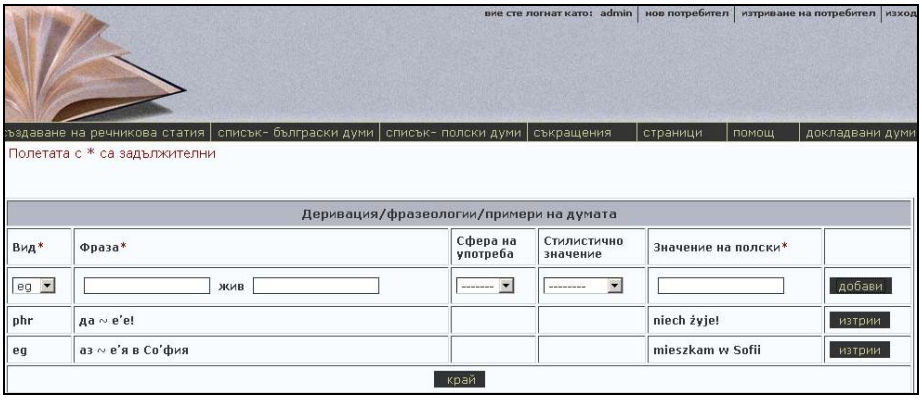

Fig. 4. Administrative module – adding examples, derivations and phrases for the verb "**живея**"

4.2 End-User Module of the Bulgarian-Polish Online Dictionary

The end-user module is intended to and generally accessible by the casual users. It provides the functionality to search for meanings of words in Bulgarian to Polish and vice versa. This module has a relatively simple structure: two mirror versions "Bulgarian" and "Polish", and the following sections: "Dictionary", "Project" and "Support". The section "Support" provides link from casual users to the dictionary author-

ized administrator: the casual user can report for a word currently missing in the dictionary or to warn about errors or gaps in the dictionary entries.

If some user doesn't dispose with keyboard supporting the Cyrillic or/and the Polish diacritics letters virtual keyboards are implemented for both alphabets, so the letters can be selected with the help of the mouse.

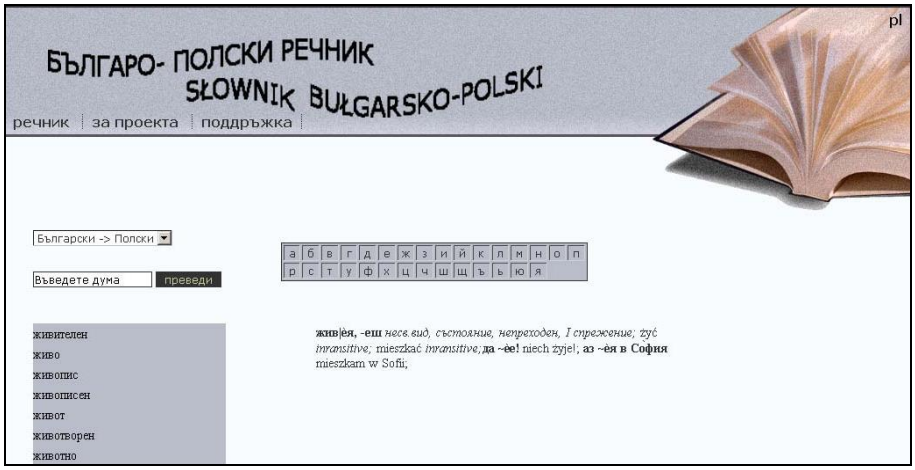

Fig. 5. End-user module – translations of the Bulgarian word "**живея**" to Polish

When casual users search translations of Polish word to Bulgarian only the Bulgarian headword will be visualized:

Fig. 6. End-user module – translation of the Polish word "**mieszkać**" to Bulgarian

5 Conclusion

In this paper the first Bulgarian-Polish online dictionary is briefly described. The dictionary is still at an experimental stage and is intended for research purposes, but it will be applicable in the daily life for educational and translation purposes. Extending of the dictionary is feasible [3]. The established Bulgarian-Polish parallel corpus, which contains more than 3 million words, can provide a good base of data for bilingual dictionary creation – many examples of the usage of the words from the corpus in a wide context could be extracted. The first Bulgarian-Polish online dictionary has the potential to develop, to enrich and become widely available and usable tool [2].

The recent version of the dictionary works optimally on Internet Explorer 6.0 + (Windows), Firefox 2.0.1 + (Windows, Linux). The resolution on the site is 1024/768 pixels. The designed modules are functional and comfortable for use.

References

1. Andreychin, L., Georgiev, L., Ilchev, St., Kostov, N., Lekov, I., Stoikov, St., Todorov, Tsv.: Bulgarian Explanatory Dictionary. 4th revised edition. Dimitar G. Popov editor. Sofia, Nauka i Izkuvstvo Publishing House, 1–1093 (In Bulgarian) (1994)
2. Dimitrova, L., Dutsova, R., Panova, R.: Survey on Current State of Bulgarian-Polish Online Dictionary. In: International Workshop "Language Technologies for Digital Humanities and Cultural Heritage" within International Conference RANLP'2011, Hissar, Bulgaria. 43–50 (2011)
3. Dimitrova, L., Koseska, V., Dutsova, R., Panova, R.: Bulgarian-Polish online Dictionary – Design and Development. In: Koseska, Dimitrova, Roszko (Eds.), Representing Semantics in Digital Lexicography. MONDILEX Fourth Open Workshop, Warsaw, SOW, 76–88 (2009)
4. Dimitrova, L., Panova, R., Dutsova, R.: Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop, Bratislava, Slovak Republic, Tribun, Brno, 36–47 (2009)
5. Dimitrova, L., Pavlov, R., Simov, K.: The Bulgarian Dictionary in Multilungual Data Bases. Cybernetics and Information Technologies. Sofia, 2 (2): 12–15 (2002)
6. Erjavec, T., Evans, R., Ide, N., Kilgarriff, A.: The Concede model for lexical databases. In: 2nd International Conference on Language Resources and Evaluation, LREC'00, Athens, ELRA (2000)
7. Ide, N., Véronis, J.: Encoding dictionaries. In: Ide, N., Veronis, J. (Eds.) The Text Encoding Initiative: Background and Context. Dordrecht: Kluwer Academic Publishers, 167–179 (1995)
8. Sławski F.: Podręczny słownik Bułgarsko-Polski z suplementem. 2nd edition, Warszawa, Polska (1987)