Information Technologies for the Preservation of Language Heritage

Ludmila Dimitrova[1], Ralitsa Dutsova[2], Rumiana Panova[2]

[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
ludmila@cc.bas.bg
[2]Veliko Tărnovo University & IMI-BAS Master Program, Sofia, Bulgaria
r.dutsova@yahoo.com, rumiana.panova@gmail.com

Abstract. In this paper we try to present how information technologies as tools for the creation of digital bilingual dictionaries can help the preservation of natural languages. Natural languages are an outstanding part of human cultural values and for that reason they should be preserved as part of the world cultural heritage. We describe our work on the bilingual lexical database supporting the Bulgarian-Polish Online dictionary. The main software tools for the web-presentation of the dictionary are shortly described. We focus our special attention on the presentation of verbs, the richest from a specific characteristics viewpoint linguistic category in Bulgarian.

Keywords: annotation tag, digital dictionary, dictionary entry, information technologies, lexical and relational data bases

1 Introduction

The successful implementation of recent developments in information technology offers numerous tools with a wide range of applications, especially for natural language processing. Digital dictionaries are great repositories of data. Actually, every dictionary contains a large amount of language data, but a digital one contains incomparably more because it is a dynamic collection of dictionary entries and has the potential for infinite growth: new entries can be added without limitation. Digital dictionaries are one of the well-known tools for applications in the social sciences and digital humanities and therefore could be used successfully for the preservation of language heritage. All kinds of digital data are now accessible from remote computers via the Internet.

Online dictionaries published in Internet are accessible to every user through a URL-address. In order to use this kind of dictionary, the user does not need any necessary hardware on the local computer or any installation of necessary software. The only condition is that the user's computer be equipped with a web browser. This is why online dictionaries are so easy to distribute. A programmer of such applications can easily and promptly correct any potential shortcoming that arises, since the application is installed on a web-server. Another advantage of online

dictionaries is that changes in their content such as deletion or addition of new dictionary entries are achieved through the so-called Content Management System.

As disadvantages of online dictionaries, we can point to the necessity for the user to have Internet access, the speed of database searches depends on the number of users using the web-based software simultaneously; and if the web-server stops working, then no single user can use the dictionary.

The building of bilingual digital dictionaries is a complex and difficult process, due to the lack of enough formal models that adequately reflect the specific linguistic features of a given natural language.

The main design goals of the Bulgarian-Polish online dictionary are: to be a general purpose dictionary, to be oriented to the casual user and be available by open access via the Net, to include within entries links to other entries, and to be easy updated or edited online (through the correction of eventual mistakes, or the addition of new entries or new information about headwords). For the realization of these purposes we need to develop a bilingual lexical database (LDB) supporting such a web-based dictionary and ensuring a good search system. Besides, whenever possible the LDB should automatically generate a new (whether a single or multiple) structure/s of entries for the Polish-Bulgarian dictionary using the appropriate information from a Bulgarian-Polish entry.

2 Bulgarian-Polish Lexical Database

The LDB of the Bulgarian-Polish online dictionary follows the CONCEDE model for dictionaries encoding with some extensions and modifications. The project CONCEDE[1] has built lexical databases in a general-purpose document-interchange format, for the six Central and East European languages: 2500-headword lexical databases for Bulgarian, Czech, Estonian, Hungarian, Romanian, and a 500-word one from the English-Slovene dictionary. The project has produced lexical resources that respect the guidelines for encoding dictionaries [9] and so are compatible with other TEI-conformant resources. The LDB model offers a standardized hierarchical tree-structure of a dictionary entry with a understandable semantics. It is formally described in [7], [8]. The first LDB for Bulgarian was developed under the CONCEDE project. It contains more than 2700 lexical entries [2] prepared in accordance with encoding standards established by the TEI. The Bulgarian LDB is based on the Bulgarian Explanatory Dictionary [1].

Firstly, we extended the monolingual model of the CONCEDE LDB to a bilingual one. Secondly, we tried to cover more of the specific features of Bulgarian and Polish aiming more adequate presentation of these Slavic languages. CONCEDE LDBs used two types of tags: structural and content tags. All tags were encoded according to the Text Encoding Initiative (TEI) traditions.

[1] PL96-1142 INCO-COPERNICUS project CONCEDE *Consortium for Central European Dictionary Encoding*

2.1 The structural tags of the Bulgarian-Polish LDB

As CONCEDE LDB, Bulgarian-Polish LDB, uses three structural tags: **entry** (dictionary entry), **struc** (indicates separate independent part – structure – in the dictionary entry, the type of this part is determined by the sub-tag type which values are modified "sense" or a new one "function"), **alt** (alt: alternation, though generally for use in quite different contexts).

We introduced a structure of a new type "Function" in order to represent the functional homonymy of the Bulgarian words. The index of type "Function" is equal to the numbers of different grammatical functions of the specified Bulgarian entry. It marks the groups of grammatical functions that correspond to a particular part of speech of the specified Bulgarian entry.

For example, two structures of type "Function" – one structure with POS as adjective and other structure – with POS as adverb (according to the Polish lexicographic tradition we use abbreviations from Latin adi /adiectivum/ and adv /adverbium/) in the entry of the headword "**политически** /political, politically/" are created:
политически *adi.* polityczny, mający związek z polityką; *adv.* politycznie; **~а иконо'мия** ekonomia polityczna

```
<entry>
<hw>политиʹческ|и</hw>
<struc type =”Function” n="1"><pos>adi</pos>
<struc type=”Sense” n=”1”>
  <trans>polityczny</trans>
</struc>
<struc type=”Sense” n=”2”>
  <trans>mający związek z polityką</trans>
</struc>
</struc>
<struc type =”Function” n="2"><pos>adv</pos>
<struc type=”Sense” n=”1”>
  <trans>politycznie</trans>
</struc>
</struc>
<eg><q>~а иконоʹмия</q></eg>
</entry>
```

2.2 The content tags of the Bulgarian-Polish LDB

The set of content tags, includes the elements:
case (contains grammatical case information given by a dictionary for a given form), **def** (directly contains the text of the definition), **domain**, **eg** (a structure, contains an example, as given in a dictionary, and allows the tags **source** and **q**), **etym** (a structure, contains etymological information and allows the tags **lang** and **m**, as given in a dictionary), **gen** (identifies the morphological gender of a lexical item, as given in

the dictionary), **geo** (geographic area), **gram** (contains grammatical information relating to a word other than gender, number, case, person, tense, mood, itype, as these all have their own element, for example, *perfect aspect* or *progressive aspect*), **hw** (the headword; used for alphabetization and indexing, access), **itype** (indicates the inflectional class associated with a lexical item, as given in a dictionary), **lang** (language; for use in etymologies (etym)), **m** (indicates a grammatical morpheme in the context of etymology), **mood** (contains information about the grammatical mood of verbs, as given in a dictionary), **number** (indicates grammatical number associated with a form, as given in a dictionary), **orth** (gives the orthographic form of a dictionary headword), **person** (indicates grammatical person associated with a form, as given in a dictionary), **pos** (indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.)), **q** (contains a quotation or apparent quotation), **register** (register, for type attribute on usg tag), **source** (bibliographic source for a quotation), **subc** (contains sub-categorization information (for verbs: transitive/intransitive, for numerals: countable/non-count, etc.)), **time** (temporal, historical era, for example, “archaic”, “old”, etc.), **tns** (indicates the grammatical tense associated with a given inflected form in a dictionary), **trans** (contains translation text and related information, so may contain any of the basetags; the principle is that everything under trans relates to the target language), **usg** (contains usage information in a dictionary entry, other than time, domain, register (as these all have their own element), like “dialect”, “folk”, “colloquialism”, etc.), **xr** (uses to indicate a cross reference with the pointer).

For each group of synonym Polish translations of a given Bulgarian word, a corresponding structure of type “Sense” is created. The Polish translation of Bulgarian headword appears in the entry in structures of type "Sense" indexing by the numbers of synonymous group of translations, for example:

бла’г *adi.* łagodny, błogi, kojący

```
<entry><hw> бла’г </hw>
  <pos>adi</pos>
<struc type=”Sense” n=”1”>
  <trans>łagodny</trans/
</struc>
<struc type=”Sense” n=”2”>
  <trans>błogi</trans/
</struc>
<struc type=”Sense” n=”3”>
  <trans>kojący</trans/
</struc>
</entry>
```

3 Digital Presentations of Some Specific Features of Bulgarian

The structure and content tags of the designed structural unit should fully meet international standards so that the LDB and the electronic dictionaries are compatible with language resources created in other projects and for other languages. Detailed information about the presentation of language specifications of Bulgarian and Polish as dictionary entry classifiers could be found in [3], [4].

Structure of a dictionary entry:

- Headword
- Formal Features – phonetics, grammar, morphology, syntax, etymology, style
- Semantic information
- Quotations
- Additional information:
  1. Derivatives
  2. Phrases
  3. Examples - phrasal and sentence usages, illustrations

The CONCEDE model [7] is used in the development of LDB for 6 European languages – Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The first LDB for Bulgarian was created in the above framework and contains more than 2700 lexical units from the "Bulgarian Explanatory Dictionary" [1].

3.1 Realization of homonyms

The meanings of homonyms are entered in the digital dictionary as different database records. On the word-entry page, there is a field where the user must specify a homonym index - a number which shows the order of the meanings. For the representation of the homonym it is necessary to fill in the value of the attribute n (homonym index) in the tag <entry>:

```
<entry n="1">
<hw>я'с|ен</hw>
.............
<entry n="2">
<hw>я'с|ен</hw>
.............
```

These entries (hw **ясен** /*clear, serene*/ in entry n="1" and hw **ясен** /*ash-tree*/ in entry n="2") will be shown on the screen as follows:

I **я'с|ен, -на, -но** *adi.* jasny; **като' гръм от ~но небе'** *pot.* jak grom z jasnego nieba; **~ен по'черк** wyraźny charakter pisma

II **я'с|ен, -и** *m bot.* jesion *m* (Fraxinus excelsior)

3.2 Presentation of Bulgarian Verbs

Nearly all verb forms and their respective grammatical meanings which once existed in the old-Bulgarian language have been preserved in contemporary new Bulgarian, in contrast with the rest of the Slavic languages which have significantly simplified their verb systems. The Bulgarian verb has a very well developed system for the expression of the "tense" category - there are forms to express 9 different verb tenses. The verb also supports expression of the following grammatical categories: person, number, voice, aspect, tense, and mode.

Depending on particularities of their lexical meaning, Bulgarian verbs are transitive (allow a direct subject - the action is transferred from the subject to another object) and intransitive (the action is not transferred to an object). Transitive verbs predominantly indicate concrete actions (write, build, draw) or perceptions and feelings (see, hear, notice), while intransitive verbs indicate movement and position (run, walk, swim), a physical or mental condition (be sick, be silent, sleep), or the change in condition of a person or an object (lose weight, gain weight, dry).
In order to adequately present the Bulgarian verb in the LDB and respectively in the dictionary, we introduced several new tags.

The content tag <**subc>** that contains sub-categorization information is very useful for the presentation of specific information pertinent to Bulgarian verbs, namely information about their transitivity/intransitivity //(**преходен/непреходен глагол** in Bulgarian)//. New information is also added: in the tag <**gram**> - for the aspect of verbs for perfect aspect and progressive aspect //(**свършен/несвършен вид** in Bulgarian//; in the tag <**conjugation**> - for the conjugation (there are 3 types of conjugation of Bulgarian verbs), and in the tag <**semantic**> - for the expression of active indication of the verb action (event or state//**събитие** или **състояние** in Bulgarian//).

The following example shows the presentation of the Bulgarian verb **боря/боря се** //*fight, struggle*// in the Bulgarian-Polish LDB:

```
<entry>
  <hw>бо'р|я</hw>
    <conjugation>
      <orth>-иш</orth>
      <type>2</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>niepokoić</trans>
</struc>
```

```
<struc type="Sense" n="2">
  <trans>męczyć</trans>
</struc>
<struc type="Derivation" n="1">
  <orth>~ся ce</orth>
  <trans>borykać się, walczyć, zmagać się</trans>
</struc>
</entry>
```

4 Relational Database (RDB)

The model of the relational database is based on the researched lexical entries. As the number of these lexical entries is limited we assume that the relational database is experimental and can be improved with the increasing number of examined lexical entries.

In the design of the relational database we also try to provide the opportunity for translation from Polish to Bulgarian. In that case, however, when one wants to automatically obtain a new Polish-Bulgarian dictionary entry from a given Bulgarian-Polish one, as well as translate in both directions, one encounters a difficult problem. This problem is related to the fact that in each language lexical forms have more than one meaning, and these meanings do not overlap in two-way translation. That is why translation will only be made from the main senses of the Bulgarian headwords. Phrases and examples cannot provide synonymous meanings, and therefore they will not be used to translate from Polish to Bulgarian. Human editing is obligatory in that case.

The current model of the relational database is represented on Figure 1. Detailed information on it can be found in the Appendix: Table 1 and Table 2.

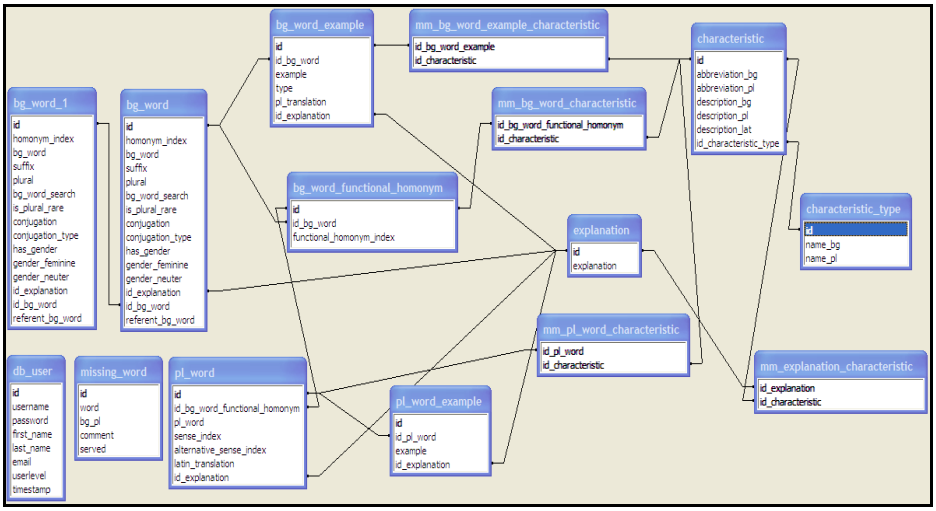

Fig. 1. Relational database upon the lexical database of the BG-PL-BG Dictionary

5 LDB into RDB transformation

For transforming the lexical database into the relational database we created an XML parser. We fixed on the DOM (Document Object Model) technology. With this technology the whole document is read and a DOM tree is constructed. This tree represents a hierarchy of nodes and each node is an object in the XML document. A random access to the nodes of the DOM tree is possible. All the embedded tags and attributes of the current node can be accessed by random too [5].

For that reason the DOM technology is chosen instead of the alternative SAX technology which cannot process complex and embedded searches. The disadvantage of the DOM technology is the more memory required when reading large XML documents than in SAX technology.

The DOM parser for transforming the LDB of the Bulgarian–Polish Dictionary into RDB is programmed in Java. In this way it can be run on different platforms no matter of the architecture or the operating system.

6 Online Dictionary – Brief Description

The recent web-based application is experimental, and the structure of the text fields is not final [6]. Future examination of the Bulgaria-Polish dictionary is a precondition for changes in the database and web application. The technologies used for the implementation of the web-based application are Apache, MySQL, PHP and JavaScript. We use free technologies originally designed for developing dynamic web pages with many functionalities.

The web-based application consists of administrator and end-user modules. The administrator module is used to fill in the database and to offer user-friendly interface to the person who will be responsible for the dictionary management. The administrator module is intended for the person updating the dictionary. The access to the administrative module will be possible only for authorized users. After logging in to the system, the user has the ability to enter a new word, to search for Bulgarian or Polish words in the database, to enter new abbreviations or to edit/delete an existing one. The interface of the web-based application is bilingual. The user can choose the input language (Bulgarian or Polish), and there is a possibility to search for a translation in both directions: Bulgarian to Polish or Polish to Bulgarian. The translation from Bulgarian to Polish will display the whole information existing in the dictionary entry, but the translation from Polish to Bulgarian will be formed only from the main senses of the Bulgarian headwords. In the end-user module there is a contact form where the casual user can report for words currently missing in the dictionary or to warn about errors or gaps in the dictionary entry. The example in Figure 2 shows how a Bulgarian verb will be entered in the LDB through the administrative module of the web application (especially the information about its transitivity, semantic features and conjugation type). Furthermore, the examples in Figure 3 and Figure 4 show how this information will be displayed on the screen to the end-user (in a translation from Bulgarian to Polish and from Polish to Bulgarian).

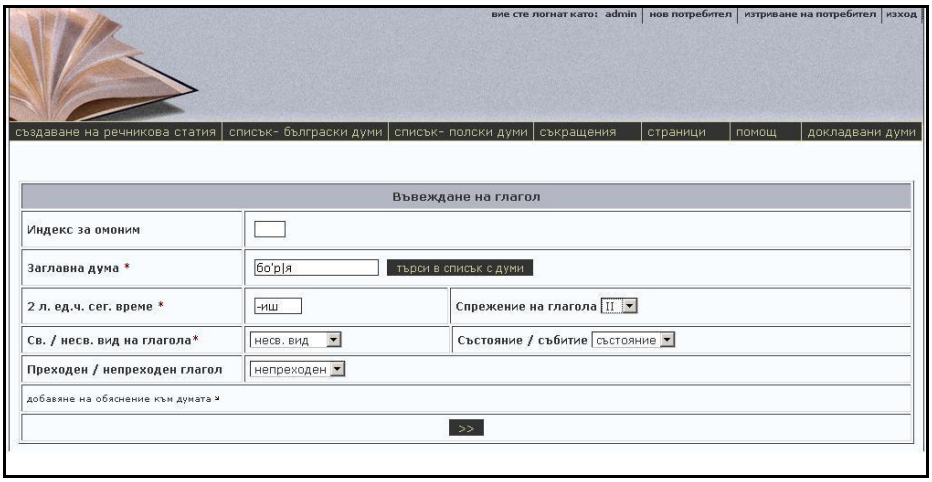

Fig. 2. Entering of a Bulgarian verb

Fig. 3. End-user screen: translation from Bulgarian

Fig. 4. End-user screen: translation from Polish

7 Conclusion

In this paper we present the recent development of the LDB supporting the digital online Bulgarian-Polish dictionary.

The bi- and multilingual digital LDB and dictionaries are large repositories of linguistics data and are one of the well-known tools for applications in the social sciences and digital humanities. They build bridges between people and cultures, and support the world's cultural heritage. Besides, the digital bilingual dictionaries in Slavic languages will be widely applicable to the contrastive studies of these languages, in system for human and machine translation, as well as in education.

References

1. Andreychin, L. et al.: Bulgarian Explanatory Dictionary. /Dictionary of the Bulgarian Language. 4th revised edition, prepared by Dimitar G. Popov/. Nauka i Izkuvstvo Publishing House, Sofia (1994) (in Bulgarian)
2. Dimitrova, L., Pavlov, R., Simov, K.: The Bulgarian Dictionary in Multilingual Data Bases. J. Cybernetics and Information Technologies. v. 2, n. 2, 12–15 (2002)
3. Dimitrova, L., Koseska-Toszewa, V., Satoła-Staśkowiak, J.: Towards a Unification of the Classifiers in Dictionary Entry. In: Garabík (ed.) Metalanguage and Encoding Scheme Design for Digital Lexicography. pp. 48–58. Bratislava, (2009)
4. Dimitrova, L., Koseska, V.: Classifiers and Digital Dictionaries. J. Cognitive Studies/Études Cognitives. 9, 117–131 (2009)
5. Dimitrova, L., Panova, R., Dutsova, R.: Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabík (ed.) Metalanguage and Encoding scheme Design for Digital Lexicography. pp. 36–47. Bratislava (2009)
6. 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. pp. 76–88. SOW, Warsaw, (2009)
7. 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)
8. Erjavec, T., Evans, R., Ide, N. Kilgarriff, A.: From Machine Readable Dictionaries to Lexical Databases: the Concede Experience. In: 7th International Conference on Computational Lexicography, COMPLEX'03, Budapest, Hungary (2003)
9. Ide, N., Véronis, J.: Encoding dictionaries. In: Ide, N., Veronis, J. (eds.) The Text Encoding Initiative: Background and Context. pp. 167–179. Dordrecht: Kluwer Academic Publishers (1995)

Appendix

Table 1. bg_word – Bulgarian headwords

Field	**Comments**
id	Id
homonym_index	Index of the homonym (if null, no homonym exists)
bg_word	Bulgarian headword
suffix	Suffix
plural	Plural form for a noun
is_plural_rare	Frequency of usage of the plural form for a noun (null – normal, 0 - often, 1 – rare)
conjugation	Conjugation form for a verb (2 p., present)
conjugation_type	Type of conjugation for a verb (1, 2 or 3)
has_gender	Whether a noun has feminine and neuter gender
gender_feminine	Feminine gender form for an adjective
gender_neuter	Neuter gender form for an adjective
id_explanation	Foreign key to "explanation"
id_bg_word	Id of the referent Bulgarian word
referent_bg_word	Referent Bulgarian word

Table 2. pl_word – Polish headwords

Field	**Comments**
id	Id
id_bg_word_functional_homonym	Foreign key to "bg_word functional homonym"
pl_word	Polish headword
sense_index	Index of the sense
alternative_sense_index	Index of the alternative sense
latin_translation	Latin translation of the word
id_explanation	Foreign key to "explanation"