Survey on Current State of Bulgarian-Polish Online Dictionary

Ludmila Dimitrova  
IMI-BAS  
Acad. G. Bonchev St bl. 8  
1113 Sofia, Bulgaria  
ludmila@cc.bas.bg  

Ralitsa Dutsova  
Veliko Târnovo University and  
IMI-BAS Master Program  
Sofia, Bulgaria  
r.dutsova@yahoo.com  

Rumiana Panova  
Veliko Târnovo University and  
IMI-BAS Master Program  
Sofia, Bulgaria  
rumiana.panova@gmail.com  

Abstract

The dictionaries are among the well-known tools for applications in everyday life, education, sciences, humanities, and human communication. The recent developments of information technologies contribute to the design and creation of new software tools with a wide range of applications, especially for natural language processing. The paper presents an online bilingual dictionary as a technological tool for applications in digital humanities, and describes the structure and content of the bilingual Lexical Database supporting a Bulgarian-Polish online dictionary. It focuses especially on the presentation of verbs, which form the richest from a specific characteristics viewpoint linguistic category in Bulgarian. The main software modules for web-presentation of this digital dictionary are also shortly described.

1 Introduction

Recent developments in information technology have been successfully implemented in natural language processing, producing numerous tools with a wide range of applications. Digital dictionaries, being large-scale data repositories, are a popular tool for applications in everyday life, education, social sciences and digital humanities. 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.

All kinds of digital data are now accessible from remote computers via the Net. Online dictionaries freely 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 and use. A programmer of such software 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 the possibility of changes in their content such as deletion or addition of new dictionary entries.

The first Bulgarian-Polish online dictionary was designed to be a general purpose dictionary oriented to the casual user and be available by open access via the Net. Authorized users can be provided with the ability to include within entries links to other entries, and to update or edit easily online (through the correction of eventual mistakes, or the addition of new entries or new information about headwords).

For the realization of these purposes a bilingual lexical database (LDB) supporting such a web-based dictionary and ensuring a good search system has to be developed. 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.

The building of bilingual digital dictionaries is a complex and difficult process, due to the scarcity of formal models that adequately reflect the specific linguistic features of a given natural language. The lexical database has been chosen as a technological platform to support the Bulgarian-Polish online dictionary for free presentation and open access in the Internet.

The design of the Bulgarian-Polish LDB follows the CONCEDE model for dictionary encoding with some extensions and modifications. The project CONCEDE\(^1\) has built lexical databases

\(^1\) CONCEDE INCO-Copernicus project no. PL96-1142

Consortium for Central European Dictionary Encoding
The project has produced lexical resources that respect the guidelines for encoding dictionaries (Ike and Véronis, 1995) 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 (Erjavec et al., 2000, 2003).

The first LDB for Bulgarian was developed under the CONCEDE project. It contains more than 2700 lexical entries (Dimitrova et al., 2002) prepared in accordance with encoding standards established by the TEI (Text Encoding Initiative). The Bulgarian LDB is based on the Bulgarian Explanatory Dictionary (Andreychin et al. 1994).

2 Bulgarian-Polish Lexical Database

The monolingual CONCEDE LDBs used two types of tags encoded according to the TEI: structural tags and content tags.

The bilingual dictionary needs a bilingual LDB. To meet the set goal, the CONCEDE model had to modified first and the monolingual LDB had to be extended to a bilingual one. Second, new tags were added to the bilingual LDS to cover more of the specific features of Bulgarian and Polish aiming more adequate presentation of both Slavic languages.

The brief description of the Bulgarian-Polish LDB tag set follows.

2.1 The structural tags of the Bulgarian-Polish LDB

Just like the CONCEDE LDB, the Bulgarian-Polish LDB uses three structural tags: entry, struc, alt. Each structural tag plays a corresponding role as follows:

alt: shows an alternation, nevertheless generally used in quite different contexts

entry: indicates main units of the BDB – dictionary entry

struc: indicates separate independent part (structure) in the dictionary entry. The type of this part is determined by the sub-tag type. The values of the type are modified “Sense” or a new one “Function”.

The structure of a new type "Function" is introduced in order to represent different grammatical functions of some Bulgarian words, because the translation correspondences in Polish are different. The index of type "Function" counts the groups of grammatical functions that correspond to a particular part of speech of the specified Bulgarian word.

For example, for the following entry

приятелски adi. przyjacielski; adv. po przyjacielsku

two structures of type “Function” are created. The first structure (index n=”1”) represents the grammatical function adjective of a headword “приятелски”/friendly/ (part of speech is adjective), and the second structure (index n=”2”) represents the grammatical function adverb (part of speech is adverb):

```
<entry>
<hw>приятелски</hw>
<struc type="Function" n="1">
  <pos>adi</pos>
  <struc type="Sense" n="1">
    <trans>przyjacielski</trans>
  </struc>
</struc>

<struc type="Function" n="2">
  <pos>adv</pos>
  <struc type="Sense" n="1">
    <trans>po przyjacielsku</trans>
  </struc>
</struc>
</entry>
```

Note: Latin abbreviations adi /adjectivum/ for adjective and adv /adverbium/ for adverb are used.

2.2 The content tags of the Bulgarian-Polish LDB

The set of content tags includes the following elements:

- case: contains grammatical case information given by a dictionary for a given form
- conjugation: a new tag, contains information about the conjugation of the Bulgarian verbs
- def: directly contains the text of the definition
- domain: 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, per-
son, tense, mood, itype, as these all have their own element; for example, for aspect – perfect aspect (p.) and progressive aspect (i.)

hw: the headword; used for alphabetization and indexing

itype: indicates the inflectional class associated with a lexical item, as given in a dictionary

lang: language; for use in etymologies (in 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

semantic: a new tag, containing the active indication of the verb action (event or state)

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: new tag 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:

```xml
<entry>
  <hw>галерия</hw>
  ...
  <struc type="Sense" n="1">
    <trans>galeria obrazów</trans>
  </struc>
  <struc type="Sense" n="2">
    <usg type="register">górn.</usg>
    <trans>chodnik</trans>
  </struc>
</entry>
```

3 Digital Presentation 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.

Let us introduce some notation used in the lexical database. The symbol “’” is used to mark the accent of the Bulgarian words, and the symbol “|” is used to separate the variable part of the word from the main part.

**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

**Realization of homonyms:**

The meanings of homonyms are entered in the 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>:

```xml
<entry n="1">
  ...
</entry>
<entry n="2">
  ...
</entry>
```

**Presentation of Bulgarian Verbs:**

As expected, the richest from the viewpoint of specific characteristics is the Bulgarian verb. Traditional printed dictionaries, however, have the shortcoming that not all characteristics are coded and presented by respective classifiers (Dimitrova et al., 2009a).
To represent the Bulgarian verbs more adequately, in Bulgarian-Polish LDB new content tags were added:

- to represent the conjugation of verbs - the <conjugation> tag and the <type> tag (for the three types of conjugation),
- to represent semantic information - the <semantic> tag and the <type> tag (1 for verbs expressing “state” and 2 for verbs expressing “event”).

New information for the aspect of verbs in the tag <gram> (for perfect aspect and progressive aspect) is also added.

The content tag <subc> that contains subcategorization information is very useful for presentation of specific information of Bulgarian verbs, namely information about their transitivity/intransitivity.

The next example shows the presentation of the entry with headword повярвам /believe/ in paper Bulgarian-Polish dictionary (Sławski, 1987):

повярвам, -ш vp. uwierzyć; не могу да ~м на очи те си pot. nie mogę uwierzyć swoim oczom, nie wierzę swoim oczom

and corresponding presentation of this headword as an entry in Bulgarian-Polish LDB:

<entry>
  <hw>повярвам</hw>
  <conjugation>
    <orth>-ш</orth>
    <type>3</type>
  </conjugation>
  <semantic>
    <orth>state</orth>
    <type>1</type>
  </semantic>
  <subc>transitive</subc>
  <trans>uwierzyć</trans>
  <eg>
    <q>не могу да ~м на очи те си</q>
    <usg type="register">pot</usg>
    <trans>nie mogę uwierzyć swoim oczom, nie wierzę swoim oczom</trans>
  </eg>
</entry>

4 Relational Database of the Bulgarian-Polish Online Dictionary

The lexical database serves as the basis for designing the relational database which is the initial point for developing the Bulgarian-Polish online dictionary. Its main use is to store and search the dictionary entries.

The model of the relational database (RDB) of the Bulgarian-Polish online dictionary is based on the validated lexical entries. As the number of these lexical entries is limited, it is natural to assume that the relational database is experimental and could be improved with the increasing number of examined lexical entries.

In the design of the relational database an opportunity for translating from Polish to Bulgarian language is also provided. That translation will be made from the main senses of the Bulgarian headwords. The phrases and examples cannot provide synonymous meanings, so they will not be used for translating from Polish to Bulgarian language.

Therefore, the corresponding data for the Polish words (examples of usage, phraseology, etc.) have to be entered in the empty field in the Polish-Bulgarian dictionary entry.

The current model of the relational database is represented on Figure 1 and detailed information on it can be found in Tables 1 - 6 (see Appendix).

5 Transformation of the LDB into RDB

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, serving as a basis of the dictionary. The saved entries in the RDB can then be edited through the administrative module of the web-based application of the dictionary.

The parser implementation uses the DOM technology (Document Object Model: http://www.w3.org/DOM/DOMTR). 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 embedded tags and attributes of the current node can also be accessed at random.

For that reason the DOM technology is chosen instead of the alternative SAX (Simple API for XML) technology which cannot process complex and embedded searches. The disadvantage of the DOM technology is the higher amount of mem-
ory required when reading large XML documents compared to the 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 independent of the architecture or the operating system.

6 Online dictionary – Brief Description

The Bulgarian–Polish online dictionary is realized by the web-based application supporting by the Bulgarian-Polish LDB and RDB. This web-based application is experimental, and the structure of the text fields is not permanently determined.

The implementation of the web-based application is based on the following technologies: Apache, MySQL, PHP and JavaScript. These are free technologies originally designed for developing dynamic web pages with greater functionality.

The web-based application consists of two main software tools: administrator and end-user module (Dimitrova et al., 2009b). The administrator module serves to create the database and update the dictionary. Access to the administrative module is restricted to authorized users (so called administrators) (Table 7 in Appendix). After logging onto the system the administrator has possibilities to access the database and to enter new entries, headwords, classifications, or to edit/delete existing ones.

The web-based end-user interface is bilingual. The user can choose the input language (Bulgarian or Polish) with possibilities to search for translation in both directions Bulgarian-to-Polish, or Polish-to-Bulgarian. The Bulgarian-to-Polish translation will display the whole information existing in the dictionary entry but the opposite translation will be made only from the main senses of the Bulgarian headwords (Figure 4).

Next, an example shows how the Bulgarian verb повирам /believe/ is inserted in the data base through the administrative module of the web application (Figure 2) (especially the information about its transitivity, semantic features and conjugation type), and further, how this information is displayed on the screen to the end-user (Figure 3). (The Figures 2 – 4 are shown in the Appendix.)

7 Conclusion and Future works

The paper presents briefly the Bulgarian-Polish LDB that supports the first Bulgarian-Polish on-line dictionary. The dictionary is at an experimental stage and intended for research purposes, but it will also be widely applicable to the contrastive studies of Bulgarian and Polish, in a system for human and machine translation, as well as in education.

Future implementation will include some “search” functions with a query, where the search parameters are fixed and which as a result will extract and show to the user the relevant information from the Bulgarian-Polish LDB – dictionary entry (entries), for example, to show transitive Bulgarian verbs, or Bulgarian adjectives that serve also as adverbs.

References

Andreychin et al. 1994. Bulgarian Explanatory Dictionary /Dictionary of the Bulgarian Language. 4th revised edition, prepared by D. G. Popov/ Nauka i Izkuvstvo Publishing House, Sofia, 1994 (In Bulgarian)

Dimitrova, L., Pavlov, R., Simov, K. 2002. The Bulgarian Dictionary in Multilingual Data Bases. Journal Cybernetics and Information Technologies. 2 (2): 12-15, 2002

Dimitrova, L., Koseska-Toszewa, V., Satolastaškowiak, J. 2009a. Towards a Unification of the Classifiers in Dictionary Entry. In Garabík (Ed.), Metalanguage and Encoding Scheme Design for Digital Lexicography. Bratislava, 48-58, 2009

Dimitrova, L., Panova, R., Dutsova, R. 2009b: Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital Lexicography. Bratislava, 36-47, 2009

Erjavec, T., Evans, R., Ide, N., Kilgariff, A. 2000. The Concede model for lexical databases. In Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC’00, Athens, ELRA, 2000

Erjavec, T., Evans, R., Ide, N. Kilgariff, A. 2003. From Machine Readable Dictionaries to Lexical Databases: the Concede Experience. In Proceedings of the 7th International Conference on Computational Lexicography, COMPLEX’03, Budapest, Hungary, 2003

Ide, N., Véronis, J. 1995. Encoding dictionaries. In Ide, N., Veronis, J. (Eds.) The Text Encoding Initiative: Background and Context. Dordrecht: Kluwer Academic Publishers, 167-179, 1995

Sławski, F. 1987. Podręczny słownik Bułgarsko-Polski z suplementem. PW „Wiedza Powszechna”, Warszawa, 1987
Appendix:

Figure 1. Relational database upon the LDB of the Bulgarian-Polish online dictionary

Figure 2. Administrative panel – 1st step of inserting of a Bulgarian verb

Figure 3. Web presentation for end users - translation from Bulgarian to Polish
### Table 1. Presentation of the Bulgarian headwords

| Field               | Comments                                                                 |
|---------------------|--------------------------------------------------------------------------|
| id                  | Id                                                                        |
| id_BG_word          | Foreign key to “bg_word”                                                 |
| id_explanation      | Foreign key to “explanation”                                             |
| id_functional_homonym_index | Index of the functional homonym group                                      |

### Table 2. Functional homonymy of the Bulgarian headwords

| Field               | Comments                                                                 |
|---------------------|--------------------------------------------------------------------------|
| id                  | Id                                                                        |
| id_BG_word          | Foreign key to “bg_word”                                                 |

### Table 3. Derivations, phrases or examples of the Bulgarian headwords and their translation in Polish

| Field               | Comments                                                                 |
|---------------------|--------------------------------------------------------------------------|
| id                  | Id                                                                        |
| id_BG_word          | Foreign key to “bg_word”                                                 |
| example             | Example of the headword                                                  |
| type                | Type of the usage (1 - Derivation; 2 - Phrase; 3 - Example)              |
| pl_translation      | Polish translation                                                       |
| id_explanation      | Foreign key to “explanation”                                             |
| 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"                 |

Table 4. Presentation of the Polish headwords

| Field                              | Comments                                      |
|-----------------------------------|-----------------------------------------------|
| id                                | Id                                            |
| id_pl_word                        | Foreign key to “pl_word”                      |
| example                           | Example in Polish                             |
| id_explanation                    | Foreign key to "explanation"                 |

Table 5. Examples of the Polish headwords

| Field                              | Comments                                      |
|-----------------------------------|-----------------------------------------------|
| id                                | Id                                            |
| explanation                       | Explanation                                   |

Table 6. Explanations of the headwords, derivations, phrases and examples

| Field                              | Comments                                      |
|-----------------------------------|-----------------------------------------------|
| id                                | Id                                            |
| username                          | Username                                      |
| password                          | Password                                      |
| first_name                        | Name                                          |
| last_name                         | Family name                                   |

Table 7. Administrative users’ authorization