COGNITIVE STUDIES | ÉTUDES COGNITIVES, 10
SOW Publishing House, Warsaw 2010

LUDMILA DIMITROVA
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

MULTILINGUAL DIGITAL RESOURCES WITH BULGARIAN LANGUAGE

Abstract. The paper presents in brief Bulgarian language resources as a part of multilingual digital resources developed in the frame of some international projects, among them parallel annotated and aligned corpora, comparable corpora, morpho-syntactic specifications for corpora annotation and dictionaries encoding, lexicons, lexical databases, and electronic dictionaries.
Keywords: corpora (parallel, comparable, aligned), corpus annotation, digital dictionaries, lexical databases, morpho-syntactic specifications

1 Introduction

The first Bulgarian language TEI-compliant digital resources were developed in the Mathematical Linguistics Department of the Institute of Mathematics and Informatics (IMI) at the Bulgarian Academy of Sciences (BAS).

The Department participated in two large language engineering EC projects:

- **MULTEXT-East** *Multilingual Text Tools and Corpora for Central and Eastern European Languages*[1], 1995–1997
- **CONCEDE** ***Con***sortium for ***C***entral ***E***uropean ***D***ictionary ***E***ncoding[2], 1998–2000

A recent project in this field is the project **MONDILEX** *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources*[3], 2008–2010.

MULTEXT-East (MTE) project is a continuation of the EU *LRE MULTEXT multilingual tools and corpora* project [8], and includes six Central and Eastern European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene. The CONCEDE project successfully employs the resources developed in MTE. Our experience shows that continuity is an important prerequisite for the success of EC-financed projects.

[1] `http://aune.lpl.univ-aix.fr/projects/multext-east/`
[2] `http://www.itri.brighton.ac.uk/projects/concede/`
[3] `www.mondilex.org`

2 Bulgarian language-specific resources: morpho-syntactic specifications

The first Bulgarian language-specific resources: morpho-syntactic specifications for encoding and annotating digital corpora and lexica were developed in the EC project MULTEXT-East (MTE for short: [3]).

They contain the list of defined categories — parts of speech (POS). Each POS is encoded by a letter: noun – N, verb – V, adjective – A, pronoun – P, determiner – D, article – T, adverb – R, adposition – S, conjunction – C, numeral – M, interjection – I, residual – X, abbreviation – Y, particle – Q. A table of attribute-values is defined for each category in order to reflect the characteristic features of each so-called ***MTE languages***: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene. The characters following the POS-encoding give the values of the position-determined attributes.

The specifications define, for each part of speech, its appropriate attributes and their values, encoded by one symbol code. If a certain attribute is not appropriate for a language, for the particular combination of features, or for the word, this is marked by a hyphen in the attribute's position.

MTE use the EC project MULTEXT[4] format of lexical description — morphosyntactic description (MSD). The MSD for Bulgarian are put into accordance with Bulgarian grammar [2].

MSD consists of linear strings of characters, representing the morphosyntactic information for each word-form. The string is constructed in the following way:

- the positions of a string of characters are numbered 0, 1, 2, . . . ;
- the agreed character at position 0 encodes the corresponding part of speech: N for noun, V for verb, A for adjective, etc. ;
- each character at position 1, 2,. . . , n, encodes the value of one attribute (for nouns the attributes are: type, gender, number, case, definiteness).

For example, the MSD of the word **стената** */the wall/* is **Ncfs-y** that means POS: noun, Type: common, Gender: feminine, Number: singular, no Case, Definiteness: yes.

The proposed formalism for the MSD is not arbitrary (a MSD contains the full description of a lexical item), but has a clear and concrete aim — to be used for specific applications, incl. **corpus annotation** (the process of adding linguistic information in an electronic form to a text corpus).

The most common and important type of corpus annotation is morphosyntactic annotation (grammatical tagging or ***POS tagging***), where a **label or tag** is associated with each word in the text in order to indicate its grammatical classification. On the basis of these standard MSDs the set of corpus tags was determined.

The list of MSDs for Bulgarian contains 326 elements. To make the Bulgarian specifications more useful for annotation of corpora and automatic disambiguation of Bulgarian texts, in particular the treatment of participles, some changes were proposed [7].

[4] `www.lpl.univ-aix.fr/projects/multext/`

3 Corpora

3.1 MTE parallel annotated and aligned corpus

MTE is building an annotated multilingual corpus, composed of three major parts:

- **Parallel Corpus,**
- **Comparable Corpus**,
- **Speech Corpus** (a small one) of spoken texts in each of the six languages, comprising forty short passages of five thematically connected sentences, each spoken by several native speakers, with phonemic and orthographic transcriptions.

Multilingual parallel corpus, based on George Orwell's novel "**1984**" in the English original and the six translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene of the novel, was developed. The parallel corpus is produced as a well-structured, lemmatized, CES-corpus [9].

The texts were automatically annotated for tokenization, sentence boundaries, and part of speech annotation, using the project tools, and validated for sentence boundaries and alignment.

The alignment between the English version and translations in each of the six CEE languages produces six pair-wise alignments comprising the MTE aligned corpus. Several different software tools were used for producing such corpora.

For **Bulgarian**, the alignment was made by the Vanilla aligner — 6699 bilingual links in total.

Aligned pairs:	2-1	23	0.345190%
	1-2	36	0.540297%
	1-1	6637	99.074487%
	0-1	1	0.014970%

The next examples show excerpts of the ***Bulgarian-English aligned 1984 texts***:

1-1 Aligned sentences:

<**Obg**.1.1.3.3> **Уинстън** завъртя ключа и гласът затихна, но от това думите не станаха неразбираеми.
<**Oen**.1.1.3.3> **Winston** turned a switch and the voice sank somewhat, though the words were still distinguishable.

<**Obg**.1.1.3.4>Звукът от апарата (наричаше се телекран) можеше да бъде намален, но нямаше начин да се изключи напълно.
<**Oen**.1.1.3.4>The instrument (the telescreen, it was called) could be dimmed, but there was no way of shutting it off completely.

1-2 Aligned sentences:

<**Obg**.1.1.24.8>Изпитваше дълбок интерес към него не само защото беше заинтригуван от контраста на изисканите маниери с телосложението му на борец,

а много повече заради стаената увереност — или навярно не толкова увереност, колкото надежда, — че политическата правоверност на **О'Брайън** не е изрядна.
<**Oen**.1.1.25.8>He felt deeply drawn to him, and not solely because he was intrigued by the contrast between **O'Brien**'s urbane manner and his prize-fighter's physique.<**Oen**.1.1.25.9>Much more it was because of a secretly held belief — or perhaps not even a belief, merely a hope — that **O'Brien**'s political orthodoxy was not perfect.

3.2 Bulgarian-Polish corpus

The ***first Bulgarian-Polish corpus*** is being developed in the framework of the joint collaborative project "Semantics and contrastive linguistics with a focus on a bilingual electronic dictionary" between IMI-BAS and ISS-PAS, coordinated by L. Dimitrova and V. Koseska. It contains approx. 5 million words so far and consists of two corpora: parallel and comparable [4].

Bulgarian-Polish parallel corpus contains more than 3 million words mainly in works of Bulgarian and Polish authors — short stories, novels, children's literature, science fiction. A small part comprises official documents of the European Commission available through the Internet.

The corpus is composed of two parts: original Bulgarian texts with Polish translations or *vice versa* and texts in other languages translated into both Bulgarian and Polish.

The main part of texts is annotated at paragraph level that allows easy alignment at this level.

3.3 Other parallel corpora with Bulgarian

The Bulgarian (IMI-BAS) and Polish (ISS-PAS) teams are developing (currently for research purposes) the first ***Bulgarian-Polish-Lithuanian experimental parallel corpus*** [5]. The corpus contains over one million words till now. The main part of texts is annotated at paragraph level that allows easy alignment at this level.

In the framework of the joint collaborative project between IMI-BAS and ĽŠIL-Slovak Academy of Sciences a small ***Slovak-Bulgarian parallel corpus*** is currently under development only for research. A part of the texts is annotated at paragraph level and a small part of this corpus is aligned at sentence level. In the frame of such project a multilingual terminological database with Slovak and Bulgarian is developed [10].

A small ***parallel corpus with Bulgarian, Polish, Slovak, Slovene*** (**incl. English** as a hub language) texts of official documents of the European Commission available through the Internet is also currently collected.

These parallel corpora are intended for contrastive linguistic studies and applications in natural language processing, at first for LDBs and bilingual digital dictionaries development.

3.4 Comparable corpus

The Bulgarian comparable corpus includes *fiction* (texts from contemporary Bulgarian novels), *non-fiction* and *newspapers* (newspaper excerpts) subsets. The Bulgarian subset of MTE comparable corpus (approx. 200 000 words) were annotated manually. For each of the six MTE CEE languages, a comparable corpus was developed. It included two subsets of at least 100 000 words each, consisting of *fiction*, comprising a single novel or excerpts from several novels; and *newspapers*. The data was comparable across the six languages, only in terms of the number and size of texts. The entire MTE multilingual comparable corpus was prepared in CES format, manually or using ad-hoc tools.

The rest of comparable corpus contains approximately 3 million words from works of Bulgarian authors, including:

- prose: Dimitar Talev, Dimitar Dimov, Pavel Vezhinov, Yordan Radichkov,
- non-fiction: Zhelyu Zhelev's „Fascism“,
- Bulgarian translations of novels and short stories of prominent European authors.

The example below shows an excerpt from Bulgarian comparable corpus *fiction*:

```
<p>
След изиграни вече
<num>23</num>
турнира (!) и толкова много мачове (защото
<name type="person">Мануела</name>
рядко губеше в първи кръг и играеше и двойки, и смесени двойки)
дойде републиканското лично първенство за мъже и жени от
<date>22</date>
до
<date>26</date>
септември в
<name type="place">София</name>.
</p>
<quote rend="dblq">
<p>Едно дете спечели всички възможни титли</p>
</quote>
<p>
— писа
<name type="person"><abbr>Л.</abbr>Семерджиева</name> .
</p>
```

4 Lexicons

Bulgarian MTE lexicons are three, one for each corpus: *Bulgarian MTE parallel corpus, Bulgarian fiction, and Bulgarian newspapers*. They cover completely the available texts: Orwell's novel 1984, texts from contemporary Bulgarian literature

and newspaper excerpts, which form Bulgarian MTE comparable corpora. Bulgarian Orwell's lexicon is a lexical list, containing 55200 entries among them 17567 lemmata, needed for use in conjunction with the MULTEXT morphological tools.

The format of the lexicons is the same for all languages in the project. Each element of the lexicon (one entry per line) contains the following information: the inflected-form (word-form), the corresponding lemma and its standard lexical description (MSD) and has the following form:

word-form $< TAB >$ **lemma** $< TAB >$ **morphosyntactic description**

The table below shows an excerpt from one of the Bulgarian lexicon:

Word-Form	**Lemma**	**MSD**
време	=	Ncns-n
времена	време	Ncnp-n
времената	време	Ncnp-y
временна	временен	A--fs-n
временни	временен	A---p-n
временният	временен	A--ms-f
временно	=	Ra
временно	временен	A--ns-n
временното	временен	A--ns-y
времето	време	Ncns-y

5 Lexical Databases

5.1 CONCEDE LDB

The first **lexical database** (LDB) for integrated multilingual resources for Bulgarian was developed in the EC INCO Copernicus project *CONCEDE*. The lexical databases of the project CONCEDE were developed on the basis of the MTE parallel multilingual corpus. The CONCEDE project suggested a model for dictionary encoding containing a lexical database with standardized and well-understood structure and semantics.

The CONCEDE project has developed lexical databases (LDBs) in a general-purpose document-interchange format for the same six MTE CEE languages: 3000-headword lexical databases for Bulgarian, Czech, Estonian, Hungarian, Romanian, and a 500-word one from the English-Slovene dictionary.

Under the CONCEDE project was developed an ***encoding scheme for lexicographic specifications*** of the Bulgarian language, according to the standards for electronic dictionary encoding. This encoding scheme served to create the Bulgarian dictionary in the LDBs of CONCEDE. The choice of dictionary entries follows the method accepted by CONCEDE. The entries are equipped with lexicographic specifications for Bulgarian language in TEI-conformant SGML.

The electronic dictionary is based on the *Bulgarian Explanatory Dictionary*. Each entry in BDB is represented as a tree-structure.

For example, the entry with headword preposition "**без**" /*without, free of* / in the paper *Bulgarian Explanatory Dictionary* [1] is:

Без *предл.* Означава: **1**. Лишеност от нещо, липса на нещо. *Мъж без пари и къща без жени огън да ги гори.* Посл. *Без дъно крина — празен хамбар.* Посл. Излезе без шапка и горна дреха. **2**. Отделяне, откъсване, изваждане, отнемане. *Дружината без трета рота излезе на позиция. Десет без три е седем.* **Без време** — преждевременно, не навреме, много бързо. *Без време осърна, без време олете.* П.Р.Сл. **Без да** *сз.* — подчинителен обстоятелствен съюз за начин, който показва, че действие главното изречение се извършва при отсъствие на действието от подчиненото. *Заминал, без да се обади.* **Без друго** — непременно, положително, сигурно; бездруго. *Без друго ще дойда.* **Без малко***.

The corresponding entry in the LDB follows:

```
<entry>
<hw>без</hw>
<pos>предл.</pos>
<struc type="Sense" n="1">
  <def>Лишеност от нещо, липса на нещо.</def>
 <eg><q>Мъж без пари и къща без жени огън да ги гори.</q><source>Посл.
                                                    </source></eg>
  <eg><q>Без дъно крина — празен хамбар.</q><source>Посл.</source>
                                                          </eg>
  <eg><q>Излезе без шапка и горна дреха.</q></eg>
</struc>
<struc type="Sense" n="2">
  <def>Отделяне, откъсване, изваждане, отнемане.</def>
  <eg><q>Дружината без трета рота излезе на позиция.</q></eg>
  <eg><q>Десет без три е седем.</q></eg>
</struc>
<struc type="Phrases">
  <struc type="Phrase" n="1"><orth>Без време.</orth>
    <def>Преждевременно, не навреме, много бързо.</def>
    <eg><q>Без време осърна, без време олете.</q><source>П.Р.Сл.
                                                    </source></eg>
  </struc>
  <struc type="Phrase" n="2"><orth>Без да.</orth><pos>sz.</pos>
    <def>Подчинителен обстоятелствен съюз за начин, който показва, че
    действие главното изречение се извършва при отсъствие на действието
    от подчиненото.</def>
    <eg><q>Заминал, без да се обади.</q></eg>
</struc>
  <struc type="Phrase" n="3"><orth>Без друго</orth>
    <def>Непременно, положително, сигурно; бездруго.</def>
```

```
  <eg><q>Без друго ще дойда.</q></eg></struc>
  <struc type="Phrase" n="4">
    <orth>Без малко.</orth>
  </struc>
</struc>
</entry>
```

5.2 LDB supporting Bulgarian-Polish online dictionary

The formal model of the **LDB** [7] supporting the first experimental Bulgarian-Polish online dictionary is the CONCEDE model for dictionary encoding. The hierarchical structure of the dictionary entry is a well-formalised tree-structure. For a more adequate description of the Bulgarian verbs, two new tags are being introduced to represent the verb's conjugation (Bulgarian verbs are divided into 3 conjugations): **conjugation** — *a new tag* is added to represent the conjugation of verbs; its structure allows the subtag **type** for the possible types of conjugations of Bulgarian verbs. Furthermore, it is allowed to input additional information in the **gram** tag for the aspect — *perfect and progressive* (imperfect) of verbs, and in **subc** tag — for *transitivity/intransitivity* of verbs.

The selection of headwords included in this LDB is based on the Bulgarian-Polish parallel corpus. The main forms (lemmata) of the most frequent word forms in the corpus are selected. The word distribution according to parts of speech follows the CONCEDE model: open parts of speech — no more than 90 %, closed parts of speech — minimum 10% of the whole set of lemmata chosen.

Let us consider an entry of the Bulgarian-Polish LDB, whose respective dictionary entry of the Bulgarian-Polish printed dictionary is:

сп|я, -иш *vi*. spać; ~**и ми се** chce mi się spać, ogarnia mnie senność

The grammatical features of this Bulgarian verb ***спя*** /sleep/ are:

aspect — imperfect (progressive) /*несвършен вид*/, this verb is **intransitive** /*непреходен*/, its conjugation is a **II type** /*II спрежение*/.

The structure of the entry with headword ***спя*** /sleep/ in Bulgarian-Polish LDB follows:

```
<entry>
<hw>сп|я'</hw>
<pos>verb</pos>
<gram>imperfect</gram>
<conjugation><orth>-и'</orth>
        <type>II</type>
</conjugation>
<subc>intransitive</subc>
<struc type="Sense" n="1">
<trans> spać </trans>
</struc>
<struc type="Derivation" n="1">
```

<orth>~и ми се</orth>
<struc type="Sense" n="1">
<trans> chce mi się spać </trans>
<alt><trans> ogarnia mnie senność </trans></alt>
</struc>
</struc>
</entry>

6 Bilingual dictionaries

6.1 Electronic dictionary

The experimental version of the first Bulgarian-Polish electronic dictionary is prepared in WORD-format and contains approximately 20 thousand dictionary entries till now, for example:

подсигур|я, -иш *vp.* **v. подсигурявам**
подсигурява|м, -ш *vi.* zabezpieczać, zaopatrywać

This dictionary provides a part of the language material for the LDB of the web-based application that supports Bulgarian-Polish online dictionary.

6.2 Online dictionary

The Bulgarian-Polish online dictionary pursues so far experimental purposes. A LDB provides the language material for the dictionary.

The web-based application representing the Bulgarian-Polish online dictionary consists of two primary modules: administrator module and end-user module.

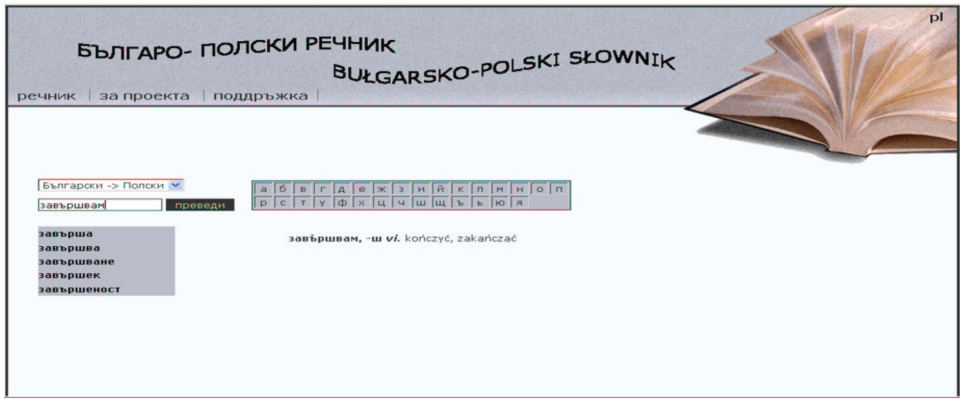
The administrator module is intended for the person updating the dictionary, and is accessible only for authorized users. The end-user module is aimed at presenting correct and up-to-date information to the user. To be convenient and easy for searching and finding the meanings of words the end-user module offers:

- An opportunity for translation from Polish to Bulgarian,
- To allow the end-user to report missing words,
- To create a user interface in both languages — Bulgarian and Polish.

For the program realization of the web-based application the IC technologies Apache, MySQL, PHP and JavaScript have been used; these are free technologies originally designed for developing dynamic web pages with a lot of functionalities. The current version of the Bulgaria-Polish online dictionary works optimally with Internet Explorer 6.0+ (Windows), and with Firefox 2.0.1+ (Windows, Linux).

The program realizing the web-based application for representation of the Bulgarian-Polish online dictionary allows the dictionary volume to be expanded by adding new words, enriching the content of the dictionary entries from the LDB by adding new examples for clarification of the meaning, etc.

The window shown below illustrates the translation of the Bulgarian verb "завършвам" /to finish/ into Polish:



The window shown below illustrates the translation of the Polish verb "kończyć" /to finish/ into Bulgarian:

In conclusion I would like to mention that the developed Bulgarian-Polish LDB may be successfully applied to the development of Bulgarian-Lithuanian LDB and of Bulgarian-Lithuanian electronic dictionary.

Acknowledgement I would like to thank all of my colleagues with whom I have worked throughout the years for the development of the Bulgarian multilingual resources: Lydia Sinapova and Kiril Simov (Bulgarian Academy of Sciences, Sofia, Bulgaria), my collaborators from the MTE and CONCEDE projects, Violetta Koseska-Toszewa (ISS-PAS), Radovan Garabík (ĽŠIL-SAS), Rumyana Panova and Ralitsa Dutsova (my students from the MSc program *Languages and Multimedia Technologies* of IMI-BAS — Veliko Tarnovo University).

References

[1] Bulgarian Explanatory Dictionary (1997). Л. Андрейчин и др. Български тълковен речник. Четвърто издание. Допълнено и преработено от Д. Попов. Издателство Наука и изкуство, София. (In Bulgarian).

[2] Bulgarian Grammar (1993). Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).

[3] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., and Tufis, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of COLING-ACL '98*. Montréal, Québec, Canada, pages 315–319.

[4] Dimitrova, L., Koseska-Toszewa, V. (2009). Bulgarian-Polish Corpus. *International Journal Cognitive Studies / Études Cognitives*. 9, SOW, Warsaw, pages 133–141, ISSN 2080-7147

[5] Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. (2009). Bulgarian-Polish-Lithuanian Corpus — Current Development. In *Proceedings of the International Workshop "Multilingual resources, technologies and evaluation for Central and Eastern European languages" in conjunction with International Conference Recent Advance in NPL'2009. Borovec, Bulgaria, 17 September 2009*. INCOMA Ltd., Bulgaria, pages 1–8, ISBN 978-954-452-008-3

[6] Dimitrova, L., Rashkov, P. (2009). A New Version for Bulgarian MULTEXT-East Morphosyntactic Specifications for Some Verbal Forms. In Shyrokov, Dimitrova (Eds. 2009), *Organization and Development of Digital Lexical Resources*. Dovira Publ. House, Kyiv, Ukraine, pages 30–37, ISBN 978-966-507-252-2

[7] Dimitrova, L., Panova, R., Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In Garabík (Ed. 2009), *Metalanguage and Encoding scheme Design for Digital Lexicography*. Bratislava, pages 36–47, ISBN 978-80-7399-745-8

[8] Ide, N., and Véronis, J. (1994). Multext (multilingual tools and corpora. In *COLING'94*, Kyoto, Japan, pages 90–96.

[9] Ide, N. (1998) Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference,* Granada, Spain, pages 463–70.

[10] Šimková, Maria, Radovan Garabík, Ludmila Dimitrova (2009). Design of a multilingual terminology database prototype. In Koseska, Dimitrova, Roszko (Eds.), *Representing Semantics in Digital Lexicography.* SOW, Warsaw, 2009, pages 123–127, ISBN 978-83-89191-87-8