

# From Electronic Corpora to Online Dictionaries (on the example of Bulgarian Language Resources)

Ludmila Dimitrova

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

**Abstract.** The paper briefly describes Bulgarian digital language resources, among them corpora, lexical databases, lexicons, and electronic dictionaries, which were developed in the Mathematical Linguistics Department at the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences (IMI-BAS) in the framework of some international projects. The first Bulgarian electronic corpora and language-specific resources were developed in the EC language technology project MULTEXT-East Multilingual Text Tools and Corpora for Central and Eastern European Languages. The first lexical database for Bulgarian was developed in the EC project CONCEDE Consortium for Central European Dictionary Encoding. These resources were developed in TEI-format, and thus they are compatible with other TEI-conformant resources. The first Bulgarian-Polish electronic corpora and dictionaries are currently developed in the frame of bilateral collaboration between IMI-BAS and ISS-PAS.

## 1 Introduction

The Department of Mathematical Linguistics at the IMI-BAS has successfully participated in the EC language technology projects MULTEXT-East and CONCEDE. The MULTEXT-East project (MTE for short: [2]), as a continuation of the project MULTEXT *Multilingual Text Tools and Corpora* [10], aims at testing and adaptation of language standards and corpus tools, developed through the MULTEXT, the development of language-specific resources for six new languages, and the extension of the annotated multilingual MULTEXT corpus. MTE developed digital language resources for six Central and East European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, as well as for English. These resources contain, for each of the six CEE languages, morphosyntactic specifications, lexica, and corpora. The corpora consist of three parts: parallel, comparable and speech-corpus. Developed in the frame of the MTE project, Bulgarian language digital resources include morphosyntactic specifications, lexicons, and corpora, incl. a parallel corpus, based on George Orwell's novel 1984, and a comparable corpus [6].

## 2 Bulgarian Language-Specific Resources: Morphosyntactic Specifications

The MTE morphosyntactic specifications have been developed on the basis of the MULTEXT specifications for Western European languages and in accordance with the EAGLES guidelines [11]. The morphosyntactic specifications have been used in the encoding of the word-form lexica of the project. They contain the list of defined categories – parts of speech (POS), each POS encoded by a letter: noun - N, verb - V,

L.Dimitrova

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 language. The specific features of each language are marked up additionally by **l.s.** 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. It should be noted that if a certain attribute is not appropriate (1) for a language, (2) for the particular combination of features, or (3) for the word, this is marked by a hyphen in the attribute's position.

The MTE use the MULTEXT format of lexical description - morphosyntactic description (MSD), which 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, etc.
- the agreed character at position 0 encodes the corresponding part of speech: N for noun, V for verb, etc. ;
- each character at position 1, 2, n, encodes the value of one attribute (for nouns the attributes are: type, person, gender, number, etc.);

*For example*, the MSD **Ncfs-** means POS: noun, Type: common, Gender: feminine, Number: singular, nocase.

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. On the basis of these standard MSDs the set of corpus tags were determined. A mapping from the morpho-syntactic information, contained in the lexical description, to a set of corpus tags is also provided, according to the MULTEXT tagging model. The list of MSDs for Bulgarian contains 326 elements.

*For example*, the MSD of the word **каптата** /the map/ is **Ncfs-y** that means POS: noun, Type: common, Gender: feminine, Number: singular, nocase, Definiteness: yes.

Some of MSDs for Bulgarian are not strictly adequate to the particular morphosyntactic properties of the respective parts-of-speech – Tense, Number, Gender, Voice, Definiteness – especially in the system of impersonal verbal forms (participles) [12]. In particular, the *present active participial* cannot possess the Tense attribute because it expresses the property/attribute independently and regardless of the tense of the main verb in the sentence, whereas the Voice attribute is also implicit from the context. New MSDs for Bulgarian participles have been proposed, bringing the morphosyntactic description in line with the grammatical characteristics of the Bulgarian, [7]. An update of the MSDs will make them more useful for annotation of corpora and automatic disambiguation of Bulgarian texts.

The Bulgarian language-specific resources also include a set of segmentation and morphological rules and data, which are necessary for use with the various annotation MULTEXT tools. Segmentation rules describe the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc. Morphological rules, needed by the morphological tools, provide exhaustive treatment of inflection and minimal derivation. The so called special tokens, required by the segmenter, includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types. Since some subtools, for example, in the segmenter require certain language-specific information in order to accomplish their tasks, each participating side

has developed a set of resource files for their language. For maximum flexibility and to retain language-independence, all such information is provided directly to the subtools via external resource files.

### 3 Corpora

MTE is building an annotated multilingual corpus, composed of three major parts: **Parallel Corpus**, **Comparable Corpus**, and a small **Speech Corpus** 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. The 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 corpus contains four parts, corresponding to the different levels of annotation: the original text of the novel, the CesDOC-encoding (SGML mark-up of the text up to the sentence-level), the CesANA-encoding (containing word-level morpho-syntactic mark-up), and the aligned versions in CesAlign-encoding (containing links to the aligned sentences). 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 a translation in each of the six CEE languages ensures six pair-wise alignments.

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

#### 1-1 Aligned sentences:

---

- <Obg.1.1.7.4>Още три сгради, подобни по външен вид и размери, бяха посети из **Лондон**.
  - <Oen.1.1.9.2>Scattered about **London** there were just three other buildings of similar appearance and size.
  - <Obg.1.1.7.5>И дотолкова се извисяваха над околните здания, че от покрива на жилищен дом **Победа** можеха да се видят и четирите едновременно.
  - <Oen.1.1.9.3>So completely did they dwarf the surrounding architecture that from the roof of **Victory Mansions** you could see all four of them simultaneously.
- 

#### 1-2 Aligned sentences:

---

- <Obg.1.1.7.3>От мястото си **Уинстън** можеше да прочете изписани с елегантни букви върхубялата фасада трите лозунга на партията: "Войната е мир""Свободата е робство""Невежеството е сила" Говореше се, че в **Министерството на истината** има три хиляди стаи над земята и съответните лабиринти отдолу.
  - <Oen.1.1.7.3>From where **Winston** stood it was just possible to read, picked out on its white face in elegant lettering, the three slogans of the **Party**: "War is peace""Freedom is slavery""Ignorance is strength."<Oen.1.1.9.1> **Ministry of Truth**, contained, it was said, three thousand rooms above ground level, and corresponding ramifications below.
-

L.Dimitrova

### ***Bulgarian MTE parallel corpus***

The Bulgarian parallel corpus contains the ***Bulgarian translation*** of Orwell's novel "Nineteen Eighty-Four", includes 86020 words (lexical items, excluding punctuation), 101173 tokens (words and punctuations); the ***CesDOC-encoding*** of the Bulgarian text of the novel (SGML mark-up of the text up to the sentence-level) includes 1322 paragraphs, 6682 sentences; the ***CesANA-encoding*** of the Bulgarian text of the novel (containing word-level morpho-syntactic mark-up), and the ***Bulgarian-English aligned texts*** - the aligned versions in ***CesAlign-encoding***, containing links to the aligned sentences (*see* examples bellow). The ***CesANA-encoding*** for Bulgarian in addition includes disambiguated lexical information for the 86020 words of the novel and undisambiguated lexical information for 156002 words. What is more – there are 156002 occurrences of MSDs in the text (Bulgarian MSD are 326) and 242022 occurrences of base or lemma of tokens (which is the total of 86020 words and 156002 occurrences of MSD). The number of occurrences of ctags is 257175. Each word-form is associated with the respective grammatical information and the corresponding lemma which form its standard lexical description.

**An example** of the ***CesANA-encoding*** of the Bulgarian text of "1984" follows:

word-form "и" /and – conjunction/, /so! oh! - interjection/

```
<tok type=WORD from='Obg.1.1.1.1\24'>  
<orth> и </orth>  
<disamb><base> и </base><ctag>CC</ctag></disamb>  
<lex><base> и </base><msd>Ccs</msd><ctag>CC</ctag></lex>  
<lex><base> и </base><msd>I-s</msd><ctag>I</ctag></lex>  
</tok>
```

### ***Bulgarian-Polish Parallel Corpus***

The first Bulgarian–Polish corpus (currently under development) is a result of the joint collaborative project "Semantics and contrastive linguistics with a focus on a bilingual electronic dictionary" between IMI—Bulgarian AS and ISS—Polish AS, coordinated by L. Dimitrova and V. Koseska. It contains a total of approximately 5 million words and comprises two corpora: parallel and comparable [3]. The first Bulgarian–Polish parallel corpus contains more than 3 million words mainly 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 corpus is developed according to the MTE model. Most texts have been annotated at paragraph level. The corpus provides samples for the experimental version of the Bulgarian–Polish digital dictionary.

In the framework of the joint collaborative project „Electronic corpora – contrastive study with focus on design of Bulgarian-Slovak digital language resources“ between IMI—BAS and IŠIL—Slovak AS, coordinated by L. Dimitrova and R. Garabik a small ***Slovak-Bulgarian parallel corpus*** is currently under development.

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.

#### **Bulgarian comparable corpora**

**Bulgarian MTE comparable corpus:** 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;
- 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 Bulgarian comparative corpus includes **Fiction** (texts from contemporary Bulgarian literature) and **Newspapers** (newspaper excerpts) subsets. The Bulgarian **Fiction** and **Newspapers** subsets were annotated manually. The data in the table below have been determined on a base of the Bulgarian fiction and Bulgarian newspapers lexica:

Part	word occurrences	distinct words	distinct MSDs in text	distinct Ctags in text
Fiction	97251	17061	313	129
Newspapers	96538	20696	295	126

The first Bulgarian electronic corpus is included in the *MTE multilingual corpus* of the MTE project (<http://nl.ijs.si/ME>), distributed on CD-ROM by *Trans-European Language Resources Infrastructure* (TELRI) Concerted Action Copernicus 1202, (<http://www.ids-mannheim.de/telri/>) for research purposes.

**Bulgarian comparable corpus in Bulgarian-Polish corpus:** This corpus contains approximately two 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.

## **4 Bulgarian Lexical Databases**

### **CONCEDE Bulgarian LDB**

The first lexical database (LDB) for Bulgarian was developed in the framework of CONCEDE project. The lexical databases of the project CONCEDE were developed on the basis of the MTE parallel multilingual corpus (so-called *Orwell* 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. The project has produced lexical resources that respect the guidelines of the Text Encoding Initiative - Dictionary Working Group (TEI-DWG), and so are compatible with other TEI-conformant resources.

L.Dimitrova

The initial word lists for selection of headwords and word frequency were obtained from the MTE parallel corpus. The selection of headwords was made after word frequency and word class (POS) were taken into account, and the number of words there were in a given word-class and word-frequency band.

In order to achieve a harmonization of the LDBs according to the principal breakdown of lemmata to POS, the CONCEDE consortium decided on the following proportion: open parts of speech (nouns, verbs, adjectives, adverbs) - no more than 90 %, closed parts of speech (numerals, pronouns, conjunctions, prepositions, particles and interjections) – minimum 10% of the whole set of lemmata chosen. Under the CONCEDE project was developed an encoding scheme for lexicographic specifications of the Bulgarian language, according to the standards for electronic dictionary encoding [5]. 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 [1]. Each entry is represented as a tree-structure. The chosen entries are divided in the following POS: noun – 33.84% of the Bulgarian sample; verb – 21.99%; adjective – 12.52%; adverb – 11.51% -- total open POS 79.86%; and numeral – 1.52%; pronoun – 5.24%; conjunction – 4.06%; preposition – 3.55%; particle – 4.40%; interjection – 1.35% -- total closed POS 20.13%. The entries in Bulgarian LDBs retain as much as possible the structure of the original paper dictionary.

The example shows the entry in the printed Bulgarian Explanatory Dictionary with the headword “**име**” //*name*//:

**име** *ср.* Отличително название на човек, животно и др. прен. Известност. *Той има голямо име.* грам. Категория думи, които означават предмети, качества, числа. *Съществително име. Прилагателно име. Числително име.* ◊ В името на предл. Въз основа на, заради. В името на закона. В името на свободата.

The corresponding entry in the Bulgarian LDBs:

```
<entry><hw>име</hw>
<gen>ср.</gen>
<struc type="Sense" n="1">
<def>Отличително название на човек, животно и др.</def></struc>
<struc type="Sense" n="2"><usg type="register">прен.</usg>
<def>Известност.</def>
<eg><q>Той има голямо име.</q></eg></struc>
<struc type="Sense" n="3"><usg type="register">грам.</usg>
<def>Категория думи, които означават предмети, качества, числа.</def>
<eg><q>Съществително име.</q></eg>
<eg><q>Прилагателно име.</q></eg>
<eg><q>Числително име.</q></eg></struc>
<struc type="Phrases">
<struc type="Phrase" n="1"><orth>В името на</orth><pos>предл.</pos>
<def>Въз основа на, заради.</def>
<eg><q>В името на закона.</q></eg>
<eg><q>В името на свободата.</q></eg></struc></struc>
</entry>
```

In the final phase of the development of the CONCEDE LDBs an examination was carried out – a validation process which takes two forms, “formal validation” and “content validation”. The formal validation was a matter of ensuring that the databases were valid SGML documents and for the Bulgarian LDBs has been done by means of a validating SGML-parser. The content validation of the entries required human intervention and therefore was performed manually.

**LDB supporting Bulgarian-Polish online dictionary**

The formal model of the LDB [4] supporting the first Bulgarian-Polish dictionary is the CONCEDE model for dictionary encoding, [8]. The hierarchical structure of the dictionary entry is a tree-structure and described by 3 structural tags: **entry**, **struc**, and **alt**. The content tagset includes tags, fully describing the entry's content: the grammatical information about the headword, the translation equivalence in Polish, examples of the word's usage with translation, phrasal usage with translation (if possible) or explanation, the word's etymology (if known).

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 the dictionary's 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 POS also 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.

The representations of three Bulgarian verbal forms as entries in the LDB follow:

**подчертава**|м, -ш *vi. podkrešlać /underline/*

**подчертан** *part. podkrešlony /underlined/*

**подчерта**|я, -еш *vp. v. подчертавам*

```
<entry>
  <hw>подчерта'ва|м</hw>
  <pos>v</pos>
  <gram>i</gram>
  <subc>transitive</subc>
  <conjugation>
    <orth>-ш</orth>
    <type>III</type>
  </conjugation>
  <struc type="Sense" n="1">
    <trans>podkrešlać</trans>
  </struc>
</entry>
```

```
<entry>
  <hw>подчерта'н </hw>
  <pos>part</pos>
  <alt>
    <pos>adi</pos>
  </alt>
  <struc type="Sense" n="1">
    <trans>podkrešlony</trans>
  </struc>
</entry>
```

L.Dimitrova

```

<entry>
  <hw>подчерта'я</hw>
  <pos>v</pos>
  <gram>п</gram>
  <subc>transitive</subc>
  <conjugation>
    <orth>-еш</orth>
    <type>I</type>
  </conjugation>
  <xr>подчерта'вам</xr>
</entry>

```

Transformation of the Lexical Database to the Relational Database is carried out with the help of tables, into which the search data and indices are input. This organization allows an automatic creation of a dictionary entry for a Polish word, whenever the translation equivalence is one-to-one. Of course, the input of information about the Polish word must be done additionally.

Column / Word	подчерта'ва м	подчерта'н	подчерта'я
id	668	669	670
homonym_index			
bg_word	подчерта#ва	подчерта#н	подчерта#
suffix	м		я
bg_word_search	подчертавам	подчертан	подчертая
plural			
is_plural_rare			
conjugation	ш		еш
conjugation_type	3		1
has_gender			
gender_feminine			
gender_neuter			
id_explanation			
id_bg_word			668
referent_bg_word			подчерта#вам

Table bg\_word

id	id_bg_word	pl_word	sense_index	alternative_sense_index	latin_translation	id_explanation
1117	668	podkreślać	1	1		
1118	669	podkreślony	1	1		

Table pl\_word



id_bg_word	id_characteristic
668	17
668	57
669	44
670	18
670	57

Table mm\_bg\_word\_characteristic

id	abbreviation_bg	abbreviation_pl	description_bg	description_pl	description_lat	id_characteristic_type
17	мин. нсв.	vi	глагол от несвършен вид			5
18	мин. св.	vp	глагол от свършен вид			5
44	прич	part	причастие			6
57	прех	transitive	преходен глагол			7

Table characteristic

The LDB of the Bulgarian-Polish dictionary could be used for the design and creation of new bilingual online dictionaries in the future.

## 5 Digital Dictionaries

### *Monolingual: Bulgarian MTE lexica*

The Bulgarian MTE lexicons (three in total) cover completely the available texts: George Orwell's novel 1984, newspaper excerpts and texts from contemporary Bulgarian literature, 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 morphological analyser.

The table below represents the number of lemmata and entries, distributed according to a POS-characteristic, appeared in Orwell's novel 1984:

POS	Lemmata	Entries
<b>Nouns (total)</b>	9891	47969
<b>Nouns - masculine</b>	4180	25100
<b>Nouns - feminine</b>	4120	16493
<b>Nouns - neuter</b>	1591	6376
<b>Verbs</b>	4140	226666
<b>Adjectives</b>	2155	19397
<b>Pronouns</b>	92	110
<b>Adverbs</b>	790	790
<b>Adpositions</b>	98	98
<b>Conjunctions</b>	76	76
<b>Numerals</b>	67	67
<b>Interjections</b>	172	172
<b>Particles</b>	86	86
<b>Total</b>	<b>17567</b>	<b>295431</b>

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**

An excerpt from the Bulgarian lexicon follows:

Word-Form	Lemma	MSD
бели	беля	Ncfp-n //nuisance, mischief; bother; trouble; difficulty//
бели	беля	Vmia2s //to bleach, whiten; peel, skin; shell; hull//
бели	беля	Vmia3s //to bleach, whiten; peel, skin; shell; hull//
бели	беля	Vmip3s //to bleach, whiten; peel, skin; shell; hull//
бели	беля	Vmm-2s //to bleach, whiten; peel, skin; shell; hull//
бели	бял	A--p-n //white//
белите	беля	Ncfp-y //nuisance, mischief; bother; trouble; difficulty//
белите	бял	A--p-y //white//
белия	бял	A--ms-s //white//
белият	бял	A--ms-f //white//
белота	=	Ncfs-n //whiteness//
белота	белот	Ncms-s //belote (card game) //
белота	белот	Ncmt //belote (card game) //
белотата	белота	Ncfs-y //white, whiteness//

### ***Bilingual Bulgarian-Polish online dictionary***

The Bulgarian-Polish online dictionary is being developed for experimental purposes. A LDB provides the language material for the dictionary. For the program realization of the web-based application the 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 website resolution is 1024/768 pixels.

The web-based application consists of two primary modules: an **administrator module** and an **end-user** module.

**The administrator module** is intended for the person updating the dictionary, and access to it is limited only to authorized users. The administrative module is used to fill in the database and to offer user-friendly interface to the user who will be responsible for the word management. This module recognizes two types of users: (1) “**super administrator**”- who has all rights of adding, editing, deleting and searching for words; adding, editing and deleting users and (2) “**administrator**”- who has all rights except creating a new user and deleting an existing one.

**The administrator module** manages some main sections: a section for entering a new word (see the example below), sections for searching for Bulgarian or Polish words, a section where end-users report the missing words. The Help section serves both the administrators and the end users.



профил | нов потребител | изход | pl |

създаване на речникова статия | списък- български думи | списък- полски думи | съкращения | преводи | липсващи думи | помощ

Създаване на речникова статия

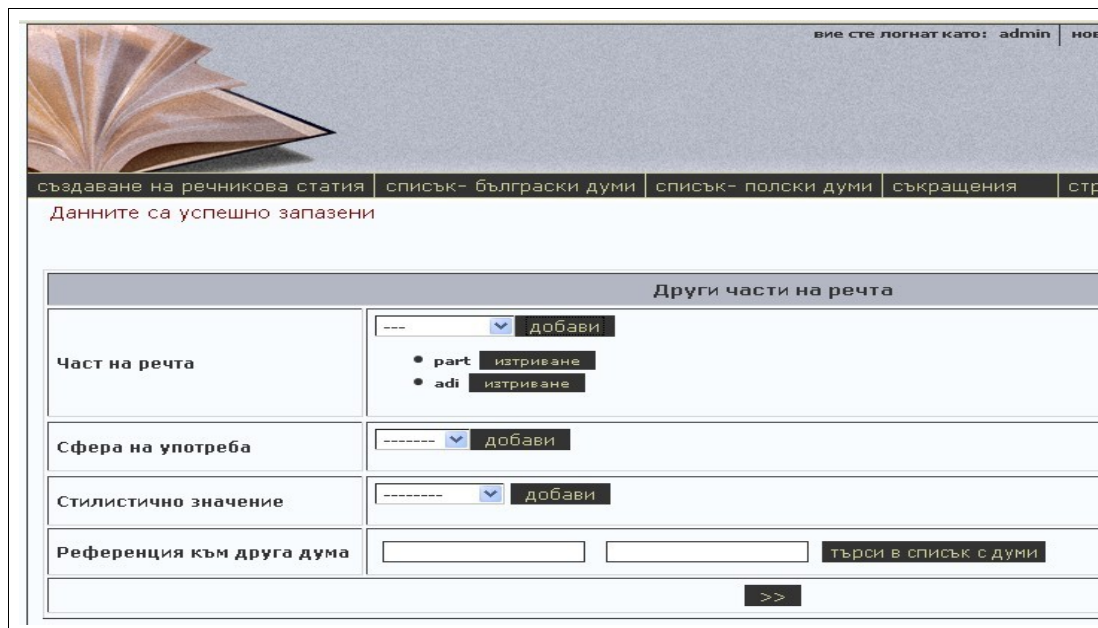
Изберете част на речта:

- съществително име
- съществително име
- прилагателно име
- глагол
- предлог

*Administrative panel - choosing the type of the word which will be added: a noun*

**The end-user module** is the module, through which the end-user accesses the information in the dictionary. The interface is bilingual, the user can choose the input language (Bulgarian or Polish) and according to his/her choice, a virtual Bulgarian or Polish keyboard is displayed. In this way the user can choose special Bulgarian or Polish characters if they are not supported by his/her own keyboard. There are three sections in this module: a section for translating a word, an information section and a section for reporting a missing word. After making a search for a word on the left site of the screen, a list of words is displayed starting from the given entry. A click on any of these words in the list visualizes the translation in the right frame. If we translate from Bulgarian to Polish, the whole information saved in the LDB is displayed. When translating from Polish to Bulgarian, only the Bulgarian headwords are visualized.

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.



вие сте логнат като: admin | noi

сздаване на речникова статия | списък- български думи | списък- полски думи | съкращения | стр

Данните са успешно запазени

**Други части на речта**

Част на речта	--- ▾ <b>добави</b>
	<ul style="list-style-type: none"><li>• part <b>изтриване</b></li><li>• adi <b>изтриване</b></li></ul>
Сфера на употреба	----- ▾ <b>добави</b>
Стилистично значение	----- ▾ <b>добави</b>
Референция към друга дума	<input type="text"/> <input type="text"/> <b>търси в списък с думи</b>

>>

*Administrative panel –2nd step of adding the participle*

Furthermore, the structures of the LDB and of the web-based application allow a replacement of the Polish translations (texts) by texts in another language Lang. Thus, the LDB and the web-based application can be useful for the development of a new bilingual Bulgarian-Lang online dictionary.

## 6 Conclusion

In this paper I briefly presented the Bulgarian language resources which were developed in the Mathematical Linguistics Department at the IMI-BAS in the framework of some international projects.

Some possible directions for future work are: bringing the morphosyntactic descriptions for verbal forms in line with the Bulgarian grammar and updating Bulgarian MSDs and lexicon for MTE resources Version 4, extending bilingual corpora, enriching bilingual LDBs with new entries and new languages, increasing the number of headword classifiers, and increasing the speed of the search module of the web-based application for representation of an online dictionary.

## Acknowledgement

I would like to thank all colleagues with whom I worked throughout the years for the development of the Bulgarian multilingual resources: Lydia Sinapova and Kiril Simov (Bulgarian Academy of Sciences, Sofia, Bulgaria), my colleagues from the MTE and CONCEDE projects, V. Koseska-Toszewa (ISS-PAS), R. Garabik (LŠIL-SAS), R. Panova and R. Dutsova (my students from the MSc program *Languages and Multimedia Technologies* of IMI-BAS – Veliko Tarnovo University).

## Bibliography

- [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, 1997. (In Bulgarian).
- [2] Dimitrova, L., T. Erjavec, N. Ide, H. Kaalep, V. Petkevič, D. Tufiş. (1998). Multext\_East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of the COLING-ACL'98*, pages 315-319, Montréal, Québec, Canada.
- [3] Dimitrova, L., V. Koseska-Toszewa. (2008). Some Problems in Multilingual Digital Dictionaries. In: *International Journal Études Cognitives*. Vol. 8. SOW, Warsaw. 2008, pages 237-254. ISSN 1641-9758.
- [4] Dimitrova, L., R. Panova, R. Dutsova. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: *Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009*, pages 36-47. ISBN 978-5-9900813-6-9.
- [5] Dimitrova, L., R. Pavlov, and K. Simov. (2002). The Bulgarian Dictionary in Multilingual Data Bases. *Cybernetics and Information Technologies*. Vol. 2, num. 2, pages 12-15.
- [6] Dimitrova, L., R. Pavlov, K. Simov, and L. Sinapova. (2005). Bulgarian MTE Corpus – Structure and Content. *Cybernetics and Information Technologies*. Vol. 5, num. 1, pages 67-73.
- [7] Dimitrova, L., P. Rashkov. (2009). A New Version for Bulgarian MULTTEXT-East Morphosyntactic Specifications for Some Verbal Forms. In: *Organisation and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009*, pages 30-37. ISBN 978-966-507-252-2.
- [8] Erjavec, E., R. Evans, N. Ide, A. Kilgarriff. (2000). The Concede model for lexical databases. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, ELRA.
- [9] Ide Nancy. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463-470, Granada, ELRA. <http://www.cs.vassar.edu/CES/>
- [10] Ide N. and J. Véronis. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing*, pages 90-96, Kyoto.
- [11] EAGLES 1996. Monachini, M. and Calzolari, N. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Report EAG-CLWG-MORPHSYN/R. <http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/>
- [12] MTE 2004. MULTTEXT-East Morphosyntactic Specifications – version 3, edition 10th May 2004.