COGNITIVE STUDIES | ÉTUDES COGNITIVES, 13: 161–170
SOW Publishing House, Warsaw 2013

DOI: 10.11649/cs.2013.010

LUDMILA DIMITROVA[1,A] & **VIOLETTA KOSESKA-TOSZEWA**[2,B]

[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia
[2]Institute of Slavic Studies, Polish Academy of Science, Warsaw
[A]ludmila@cc.bas.bg ; [B]amaz@inetia.pl

BULGARIAN-POLISH LANGUAGE RESOURCES (CURRENT STATE AND FUTURE DEVELOPMENT)

Abstract

The paper briefly reviews the first Bulgarian-Polish digital bilingual resources: corpora and dictionaries, which are currently developed under bilateral collaboration between IMI-BAS and ISS-PAS: joint research project "Semantics and contrastive linguistics with a focus on a bilingual electronic dictionary", coordinated by L. Dimitrova (IMI-BAS) and V. Koseska (ISS-PAS).
Keywords: digital bilingual resources, parallel corpus, aligned corpus, lexical data base, paper and digital dictionary, online dictionary, digital entry classifiers.

1. Introduction

The first Bulgarian-Polish digital bilingual resources (currently under development, but constantly growing with new features being added) are a result of the collaborative work under the joint research 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 (IMI-BAS) and V. Koseska (ISS-PAS). These resources include the first Bulgarian-Polish corpus and the first digital Bulgarian-Polish dictionaries.

Multilingual corpora are large repositories of natural language data with an important role in natural language processing. These digital resources are widely applicable to the contrastive studies in a multilingual context (Dimitrova, Koseska, 2012; Dimitrova, Koseska, Roszko, D. & Roszko, R., 2009a, 2009b, 2010, 2011), in a system for human and machine translation, as well as in education for the purpose of language learning or training of translators. They are a valuable multilingual dataset for language engineering research and development, especially for training of the software tools for machine translation.

What are the reasons for the development of the Bulgarian-Polish corpus? The development of a Bulgarian-Polish corpus is based on the need for research material for contrastive studies in these two languages. Bulgarian and Polish belong to the Slavic language family: Bulgarian belongs to the South-Slavic language group and Polish — to the West-Slavic language group. Linguistic and contrastive studies of these languages can be carried out based on bilingual digital resources (corpora and dictionaries). Furthermore, the first Bulgarian-Polish parallel corpus serves as a main source of vocabulary for the digital Bulgarian-Polish dictionaries.

2. First Bulgarian-Polish Corpus

The first Bulgarian-Polish corpus contains a total of approximately 5 million words and comprises two corpora: parallel and comparable (Dimitrova, Koseska, 2007, 2008, 2009a).

2.1. First Bulgarian-Polish Parallel Corpus

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 texts of official documents of the European Union available through the Internet. The corpus contains original Bulgarian texts with Polish translations or *vice versa* and texts in other languages translated into both Bulgarian and Polish.

The Bulgarian-Polish corpus is developed according to the MTE — model for multilingual corpora (COP project 106 *MULTEXT-East Multilingual Text Tools and Corpora for Central and Eastern European Languages*: (Dimitrova et al., 1998), (Dimitrova et al., 2005)).

A part of the parallel texts is annotated at paragraph level (manually or using ad-hoc tools), according to the standards of Text Encoding Initiative *TEI* and Corpus Encoding Standard *CES* (Ide, 1998) for such kind of language resources: with paragraph level (,) boundaries. The level alignment allows the drawing of a broader context in the languages.

The following table shows an excerpt from Bulgarian-Polish parallel corpus — Bulgakov's "Master and Margarita" (the text of original is also included):

BG: Кайсиевият сок вдигна обилна жълта пяна и наоколо замириса на бръснарница. Литераторите го изпиха и веднага се разхълцаха, платиха и седнаха на една пейка с лице към езерцето и с гръб към Бронная.
PL: Morelowy napój wyprodukował obfitą żółtą pianę i w powietrzu zapachniało wodą fryzjerską. Literaci wypili, natychmiast dostali czkawki, zapłacili i zasiedli na ławce zwróceni twarzami do stawu, a plecami do Bronnej.
(**RU**: Абрикосовая дала обильную желтую пену, и в воздухе запахло парикмахерской. Напившись, литераторы немедленно начали икать, расплатились и уселись на скамейке лицом к пруду и спиной к Бронной. *Часть 1, Глава 1 „Никогда не разговаривайте с неизвестными“* /Интернет-библиотека Алексея Комарова — `http://ilibrary.ru`/)

The parallel Bulgarian-Polish corpus provides a sample of the vocabulary, which is included in the Bulgarian-Polish digital dictionary.

2.2. Aligned Bulgarian-Polish Corpus

The aligned corpus is developed at the Department of Mathematical Linguistics at IMI-BAS under the supervision of L. Dimitrova. Texts of the Parallel Bulgarian-Polish corpus serve as input data. The texts of the aligned corpus are automatically annotated for language (Bulgarian or Polish) and sentence/segment boundaries. Two language-independent freely available programs are used to align Bulgarian-Polish parallel texts at the sentence level:

1. Memory Translation 2007, a computer aided tool TextAlign
   (http://mt2007-cat.ru/index.html)
2. Bitext Aligner/Converter, bitext2tmx aligner
   (http://bitext2tmx.sourceforge.net/).

These software packages have applications in computer-assisted translation. Both tools align bilingual texts without bilingual dictionaries, but still require human editing. The resulting aligned texts are similar. The aligned Bulgarian-Polish texts are manually checked for the correctness of bilingual links.

The following example presents an excerpt from the aligned at the sentence level texts of Stanislav Lem's Solaris (using TextAlign):

```
<tu tuid="0000000011">
  <tuv xml:lang="polish">
    <seg>Widziałem już seledynowy kontur jedynego wskaźnika.</seg>
  </tuv>
  <tuv xml:lang="bulgarian">
    <seg>Вече различавах светлозелените контури на универсалния указател.
    </seg>
  </tuv>
</tu>
```

Every web-based corpus is a resource combining a number of features that together make it unique and useful tool not only for language studies, but also for researchers of many fields. These features include: rich linguistic content (aligned texts in two Slavic languages for example, in this case Bulgarian and Polish), annotation (mark-up at two levels: paragraph and sentence, POS-tagging, etc.), search query (advanced possibilities for combining many search criteria), display of the search results in an intuitive and simple interface, advanced results handling (concordances, collocations, etc.).

2.3. Concordances

One of the major developments in linguistic research has come from the possibility of studying vast amounts of text through computer tools, namely through text retrieval and concordancing programs. The basic investigation procedure for querying

text corpora consists in producing multiple concordance lines, for a specified string of characters — a word, a lemma or a phrase. This means that the opportunity exists — thanks to the broader context — to study more precisely the meanings of word-forms in each language.

The following example from Bulgakov's *Master and Margarita* shows a concordance with the Bulgarian word "**литераторите**":

BG: Кайсиевият сок вдигна обилна жълта пяна и наоколо замириса на бръснарница. **Литераторите** го изпиха и веднага се разхълцаха, платиха и седнаха на една пейка с лице към езерцето и с гръб към Бронная.	**PL**: Morelowy napój wyprodukował obfitą żółtą pianę i w powietrzu zapachniało wodą fryzjerską. **Literaci** wypili, natychmiast dostali czkawki, zapłacili i zasiedli na ławce zwróceni twarzami do stawu, a plecami do Bronnej.	**RU**: Абрикосовая дала обильную желтую пену, и в воздухе запахло парикмахерской. Напившись, **литераторы** немедленно начали икать, расплатились и уселись на скамейке лицом к пруду и спиной к Бронной. (*Часть 1, Глава 1 „Никогда не разговаривайте с неизвестными“ Интернет-библиотека Алексея Комарова* `http://ilibrary.ru/`)

The Bulgarian-Polish aligned corpus will be soon represented via Internet by the Web-based software tool with a wide spectrum of features for practical applications. The corpus will be freely available for research and education on the web with an appropriate trilingual interface in Bulgarian, Polish, and English. For more detailed description of the web-application software for presentation of bilingual aligned corpora we refer to (Dimitrova, Dutsova, 2013 — in this volume).

3. Lexical Databases and Online Dictionaries

3.1. CONCEDE Bulgarian Lexical Database

The first lexical database (LDB) for Bulgarian (Dimitrova, Pavlov, Simov, 2002) was developed in the framework of INCO Copernicus project PL96-1142 *CONCEDE Consortium for Central European Dictionary Encoding*. 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 monolingual LDB with standardized and well-understood structure and semantics.

3.2. LDB supporting bilingual online dictionary with Bulgarian

The **bilingual LDB** for supporting bilingual Bulgarian-Lang2 online dictionary was designed and developed at the Department of Mathematical Linguistics at

IMI-BAS under the supervision of L. Dimitrova. The formal model of the bilingual LDB (Dimitrova, Panova, Dutsova, 2009) is the CONCEDE model for dictionary encoding (Erjavec et al., 2000). The hierarchical structure of the LDB 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 the second language Lang2, 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, some new tags are added for representing of: the Bulgarian conjugation (in total 3 conjugations) — <**conjugation**> *tag* is added to represent the conjugation of verbs; its structure allows the sub tag <**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, new tag for semantics information — <**semantic**> tag and <**type**> tag (type = 1 for verbs that mean "state", type = 2 — for "event").

The digital entry classifiers are discussed in detail in (Dimitrova, Koseska, 2007, 2008a, 2008b; Dimitrova, Koseska, Satola, 2009, 2012).

The aforementioned bilingual LDB serves as a basis for the creation of bilingual Bulgarian-Polish LDB that supports the first Bulgarian-Polish online dictionary (Dimitrova, Koseska, Panova, Dutsova, 2009).

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 (Dimitrova, Dutsova, 2012).

The representation of the Bulgarian verb **повярва|м** / *believe*/ as an entry in the LDB follows:

```
<entry>
   <hw>повя'рва|м</hw>
   <conjugation>
      <orth>-ш</orth>
      <type>3</type>
   </conjugation>
   <semantic>
      <orth>състояние</orth>
      <type>1</type>
   </semantic>
   <subc>преходен</subc>
   <pos>гл.</pos>
   <gram>св.</gram>
   <struc type="Sense" n="1">
      <trans>word in Lang2</trans>
```

```
  </struc>
  <eg><q>не мо'га да ˜м на очи'те си</q>
    <usg type="register">pot</usg>
    <trans>phrase in Lang2</trans>
  </eg>
</entry>
```

A LDB provides the language material for the dictionary. 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 Lang2 word (in this case Polish), whenever the translation equivalence is one-to-one. Of course, the input of information about the Lang2 word must be done additionally.

The implementation of the software tools realizing the web-based application for presentation of a bilingual Bulgarian-Lang2 online dictionary (Dimitrova, Dutsova, 2012) 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.

Furthermore, the structures of the LDB and of the web-based application allow a replacement of the Lang2 translations (texts) by texts in another language (in this case Polish). Thus, the LDB and the web-based application can be useful for the development of a new bilingual Bulgarian-Lang3 online dictionary.

The web-based casual/end-user interface is bilingual. The user can choose the input language (Bulgarian or Lang2) with possibilities to search for translation in both directions Bulgarian-to-Lang2, or Lang2-to-Bulgarian. The Bulgarian-to-Lang2 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

3.3. Bilingual Bulgarian-Polish online dictionary

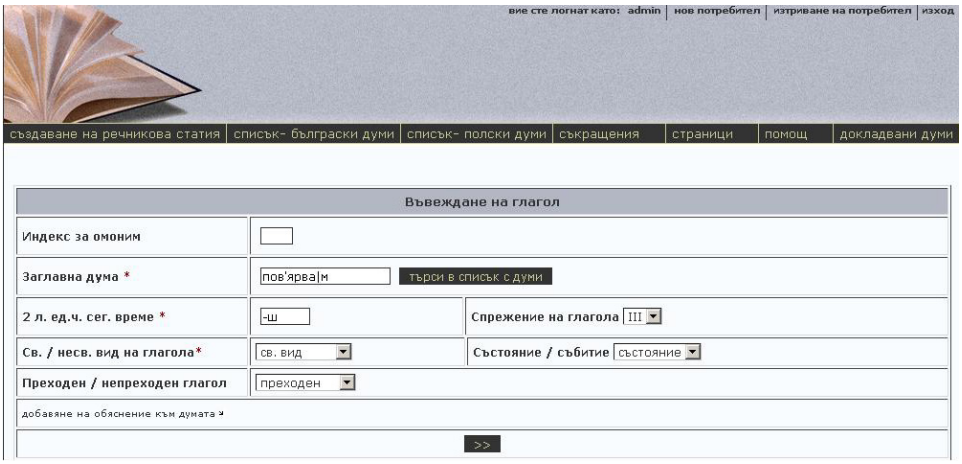
At the beginning of work on the bilateral project "Semantics and contrastive linguistics with a focus on a bilingual electronic dictionary" (2006/2007), the Bulgarian-Polish online dictionary was planned for experimental purposes only. But the implemented software tools allow preparation and presentation of the Bulgarian-Polish online dictionary as an open access tool. The first Bulgarian-Polish online dictionary is designed to be a general purpose dictionary available by open access via the Net to the casual user.

The web-based application for presentation of the Bulgarian-Lang2 online dictionary 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 a 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 **Lang2** words, a **section where** end-users report the missing words. The **Help section** serves both the administrators and the end users.

Administrative panel — choosing the type of the word which will be added



Administrative panel — 1st step of inserting of a Bulgarian verb

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 Lang2, Polish in this case) and according to his/her choice, a virtual Bulgarian or Polish keyboard is displayed. In this way the user can choose special Bulgarian or Lang2 (Polish in this case) 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.

The following examples show the web pages for the end users: translation by online dictionary of the entry with headword ***повярвам*** /believe/.

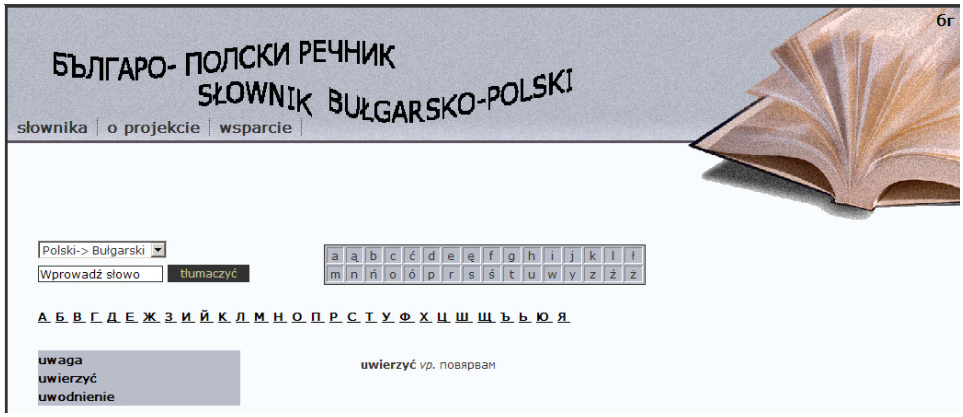
The presentation of this entry in the paper Bulgarian-Polish dictionary of F. Sławski, (Sławski, 1987) is:

повя'рва|м, -ш *vp.* uwierzyć; **не мо'га да ˜м на очи'те си** *pot.* nie mogę uwierzyć swoim oczom, nie wierzę swoim oczom.

If we translate from Bulgarian to Polish, the whole information saved in the RDB is displayed:

Web page for end users — translation of a Bulgarian verb (повярвам /believe/)

When translating from Polish to Bulgarian, only the Bulgarian headwords are visualized:



Web page for end users — translation of a Polish verb (uwierzyć /believe/)

4. Conclusion and Future Work

The paper reviews briefly the bilingual digital Bulgarian-Polish resources developed in the frame of bilateral collaboration between IMI-BAS and ISS-PAS. These resources are widely applicable to the contrastive studies in a multilingual context, in a system for human and machine translation, as well as in education.

Future implementation will include presentation on the web of some Bulgarian-Polish data, esp. of the aligned Bulgarian-Polish corpus. A web page with an appropriate trilingual interface in Bulgarian, Polish and English for easy access to the corpus is forthcoming.

References

Dimitrova, L., Dutsova, R. (2012). Implementation of the Bulgarian-Polish Online Dictionary. *Cognitive Studies | Études Cognitives*, *12*, p. 219–229, ISSN 2080-7147. DOI: 10.11649/cs.2012.015

Dimitrova, L., Dutsova, R. (2013). *Web-Application for the Presentation of Bilingual Corpora (Focusing on Bulgarian as One of the Paired Languages).* (be appear)

Dimitrova et al., (1998). L. Dimitrova, T. Erjavec, N. Ide, H. Kaalep, V. Petkevič, D. Tufiş. MultextEast: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: *Proceedings of the COLING-ACL'98, p. 315–319, Montréal, Québec, Canada.*

Dimitrova, L., Koseska, V. (2007). Digital Dictionaries — Problems and Features. In: *Proceedings of the Jubilee International Conference Mathematical and Computational Linguistics, 6 July 2007, Sofia, Bulgaria*, p. 25–34. ISBN 978-954-8986-28-1.

Dimitrova, L., Koseska-Toszewa, V. (2008a). The Significance of Entry Classifiers in Digital Dictionaries. In: L. Iomdin, L. Dimitrova (Eds.), *Lexicographic Tools and Techniques. Proceedings of MONDILEX First Open Workshop, 3–4 October 2008, Moscow*, p. 89–97. ISBN 978-5-9900813-6-9.

Dimitrova, L., Koseska-Toszewa, V. (2008b). Some Problems in Multilingual Digital Dictionaries. *Études Cognitives*, *8*, p. 237–254. ISSN 1641-9758.

Dimitrova, L., Koseska-Toszewa, V. (2009a). Bulgarian-Polish Corpus. *Cognitive Studies | Études Cognitives*, *9*, p. 133–141. ISSN 2080-7147. DOI: 10.11649/cs.2009.010

Dimitrova, L., Koseska-Toszewa, V. (2009b). Classifiers and Digital Dictionaries. *Cognitive Studies | Études cognitives*, *9*, p. 117–131. ISSN 2080-7147. DOI: 10.11649/cs.2009.009

Dimitrova, L., Koseska-Toszewa, V. (2012). Bulgarian-Polish Parallel Digital Corpus and Quantification of Time. *Cognitive Studies | Études cognitives. 12*, p. 199–208. ISSN 2080-7147. DOI: 10.11649/cs.2012.013

Dimitrova, L., Koseska-Toszewa, V., Derzhanski, I. & Roszko, R. (2009). Annotation of Parallel Corpora (on the Example of the Bulgarian-Polish Parallel Corpus). In: Shyrokov, Dimitrova (Eds. 2009), *Organisation and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop*, Kiev, 2–4 February, 2009, p. 47–54. ISBN 978-966-507-252-2.

Dimitrova, L., Koseska-Toszewa, V., Dutsova, R., Panova, R. (2009). Bulgarian-Polish online Dictionary — Design and Development. In: V. Koseska-Toszewa, L. Dimitrova, R. Roszko (Eds.), *Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open Workshop, 29 June – 1 July, Warsaw,* 2009, SOW, p. 76–88. ISBN 978-83-89191-87-8

Dimitrova et al. (2010). L. Dimitrova, V. Koseska, R. Garabík, T. Erjavec, L. Iomdin, V. Shyrokov. *Conceptual Scheme for a Research Infrastructure Supporting Resources*

in Slavic Lexicography. Sofia, Demetra Ltd. Publisher, pp. 131. ISBN 978-954-8986-33-5.

Dimitrova, Koseska, Roszko, D. & Roszko, R. (2009a). Bulgarian-Polish-Lithuanian Corpus — Problems of Development and Annotation. In: T. Erjavec (Ed.), *Research Infrastructure for Digital Lexicography. Proceedings of the MONDILEX Fifth Open Workshop within International Conference Information Society — IS 2009*, 14–15 October, 2009, Ljubljana, p. 72–86. ISSN 1581-9973/ISBN 978-961-264-012-5.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D. & Roszko, R. (2009b). Bulgarian-Polish-Lithuanian Corpus — Current Development. In: *Proceedings of the International Workshop "Multilingual resources, technologies and evaluation for Central and Eastern European languages" within International Conference RANPL'2009.* Borovec, Bulgaria, 17 September 2009. p. 1–8. ISBN 978-954-452-008-3. Available at the webpage of the Association for Computational Linguistics (ACL), `http://www.aclweb.org/anthology/W/W09/W09-4001.pdf`

Dimitrova, L., Koseska-Toszewa, V., Roszko, D. & Roszko, R. (2010). Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus). *Cognitive Studies | Études cognitives*, *10*, p. 217–240. ISSN 2080-7147. `DOI: 10.11649/cs.2010.013`

Dimitrova, L., Koseska-Toszewa, V., Roszko, D. & Roszko, R. (2011). Bulgarian-Polish-Lithuanian Corpus — Recent Progress and Application. In: *Proceedings of the Sixth International Conference NLP, Multilinguality SLOVKO'2011*, 20–22 October 2011, Modra, Slovakia, p. 30–43. ISBN 978-80-263-0049-6

Dimitrova, L., Koseska-Toszewa, V., Satoła-Staśkowiak, J. (2009). Towards a Unification of the Classifiers in Dictionary Entry. In: Garabík (Ed.), *Metalanguage and Encoding Scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, 15–16 April, 2009, Bratislava,* p. 48–58. ISBN 978-80-7399-745-8.

Dimitrova, L., Koseska-Toszewa, V., Satoła-Staśkowiak, J. (2012). Neologisms in Bilingual Digital Dictionaries (on example of Bulgarian-Polish Dictionary). *Cognitive Studies/Études Cognitives*, *12*, p. 107–114. ISSN 2080-7147. `DOI: 10.11649/cs.2012.008`

Dimitrova, L., Panova, R., Dutsova, R. (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*, p. 36–47. ISBN 978-5-9900813-6-9.

Dimitrova, L., Pavlov, R., Simov, K. (2002). The Bulgarian Dictionary in Multilungual Data Bases. In: *Cybernetics and Information Technologies*, *2*(2), p. 12–15.

Dimitrova et al. (2005). L. Dimitrova, R. Pavlov, K. Simov & L. Sinapova. Bulgarian MTE Corpus — Structure and Content. In: *Cybernetics and Information Technologies*, *5*(1), p. 67–73.

Erjavec et al. (2000). E. Erjavec, R. Evans, N. Ide, A. Kilgarriff. The Concede model for lexical databases. *In Second International Conference on Language Resources and Evaluation, LREC'00, Athens, ELRA.*

Garabík, R., Dimitrova, L., Koseska-Toszewa, V. (2011). Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). *Cognitive Studies | Études cognitives*, *11*, p. 227–239. ISSN 2080-7147. `DOI: 10.11649/cs.2011.014`

Ide, Nancy (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. *First International Conference on Language Resources and Evaluation, LREC'98*, p. 463–470, Granada, ELRA. `http://www.cs.vassar.edu/CES/`

Sławski, F. (1987). *Podręczny słownik Bułgarsko-Polski z suplementem.* PW „Wiedza Powszechna", Warszawa.