COGNITIVE STUDIES | ÉTUDES COGNITIVES, 12
SOW Publishing House, Warsaw 2012

LUDMILA DIMITROVA[1,A], **RALITSA DUTSOVA**[1,B]
[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
[A]ludmila@cc.bas.bg
[B]r.dutsova@yahoo.com

IMPLEMENTATION OF THE BULGARIAN-POLISH ONLINE DICTIONARY

Abstract

The paper describes the implementation of an online Bulgarian-Polish dictionary as a technological tool for applications in digital humanities. This bilingual digital dictionary is developed in the frame of the joint research project "Semantics and Contrastive Linguistics with a focus on a bilingual electronic dictionary" between IMI-BAS and ISS-PAS, supervised by L. Dimitrova (IMI-BAS) and V. Koseska-Toszewa (ISS-PAS). In addition, the main software tools for web-presentation of the dictionary are described briefly.
Keywords: Bulgarian, Polish, online dictionary, dictionary entry, web-presentation.

1. Introduction

Advances in information technologies for natural language processing are emerging rapidly. The recent technological developments and web-services contribute to the design and creation of new software tools with a wide range of applications, especially in the field of digital language resources. The dictionaries — together data repositories and a means for communication — are well-known tools for applications in everyday life, education, human communication, social sciences and digital humanities. 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. By that reasoning online dictionaries are widely distributed and used. A programmer-developer of such web-based software tools can easily and promptly correct any

potential shortcomings that arise, since the tools are installed on a web-server. Another advantage of online dictionaries is the opportunities for a continuous real-time update and editing, e.g. changing the content by deletion or addition of information in the dictionary entries or by adding new dictionary entries.

2. Main goal of Bulgarian-Polish online dictionary creating

The main goal of the project "Semantics and Contrastive Linguistics with a focus on a bilingual electronic dictionary" is to create an up-to-date bilingual online dictionary (Dimitrova, Koseska, Dutsova, Panova 2009).

Thus the initial tasks of the realization of this main goal were the design and development of such a dictionary by using modern and contemporary web-technologies and providing an easily used tool for managing the dictionary content, stored in a Lexical Database.

The described software package — a web-based application for presentation of the digital bilingual dictionary — is oriented to two user groups (Dimitrova, Dutsova, Panova 2011). The first user group includes so called „administrators", people that have designed, developed and managed Bulgarian-Polish online dictionary, and the second, so called "end-user" (or casual users) — people that use it.

Depending on the type of users we have allocated tasks and services in two sets: for administrators and for end-users. So the web-based application consists of two main software modules intended for "administrators" and for "end-users".

2.1 Tasks and services allocated to the "administrative" module of the online dictionary:

- To create the web-based Bulgarian-Polish dictionary, which has possibilities for presentation of the dictionary entries as a paper one, to be easy to use and does not require additional "administrator" training, to provide functionality for updating the dictionary content from the web-based software, to provide possibilities to store the information about missing words reported by the "end-users";
- To create a special kind of user — "super administrator" — who will manage the web-based application: to give access to the "administrators" (to register a new "administrator" and delete existing one), to receive the massages with information about missing words reported by the "end-user";
- To prepare a *User Manuel* that can be regularly updated.

2.2 Tasks and services allocated to the "end-user" module:

- To create a user-friendly interface in both languages — Bulgarian and Polish;
- To provide accurate and up-to-date information to "end-users";
- To ensure quick search of words in the Lexical Database (LDB) of the online dictionary;

- To provide the ability for translation from Polish into Bulgarian;
- To allow of an "end-user" to report missing words or errors and gaps in the translations of already existing dictionary entries.

3. Headwords selection procedure

The next key task of our project was the selection of the Bulgarian headwords. The applied method, statistical and linguistic at the same time, developed for CONCEDE project[1], is described in (Tufis et al. 1999). The procedure for selecting the headwords take into account word frequency, word class, and the number of words there were in a given word-class and word-frequency band. The point briefly describes a procedure, which can automatically produce Parts-of-Speech (POS) lists of any length, and then considers the manual modifications that were necessary only for the sample of the first 500 entries.

Furthermore, we adopted an approach, involving a generic sampling method for selection of headwords into the lexical database. We needed Bulgarian texts encoded as CES ANA, (Ide, Veronis 1995), which specifies for each word-form its associated lemma and grammatical information. Such texts were developed in the MULTEXT-East project[2] (Dimitrova et al. 1998). The POS composition of this sample has to reflect the corresponding distribution of the different POS in the Bulgarian MULTEXT-East corpus (Dimitrova, Pavlov, Simov 2002).

First, the corpus is divided into sequences of text, which contain 500 different lemmas of different parts of speech. In practice, the whole corpus is reduced to a sequence of <lemma, POS> pairs. Second, a counter is incremented each time a new lemma is encountered. When the counter reaches the value 500, a new text sample starts and the counter is reset to zero. This operation is repeated until the end of the corpus is reached. A statistical formula calculates the number of each POS in the sample.

This method ensures the following: the POS composition of the sample reflects the corresponding distribution of the different parts of speech in the corpus and to some extent the structural POS distribution of the language; and the number of POS lemmas chosen should not depend on the size of the corpus. The reason behind this advantage is the stylistically coherent text, from which the samples are initially taken.

Lemmas were chosen for the relevant ten grammatical categories identified in the MULTEXT-East project, according to the frequency of their occurrence in corpus.

Three frequency ranges are considered: high, medium and low. The high frequency range was assigned the interval [0.5, 1], the medium frequency range the interval [0.25, 0.5] and all the words with frequency range below 0.25 were considered in the low frequency range.

The frequency ranges were computed (for each POS) based on a normalized occurrence ranking of each word form. The normalized ranking of a lemma was

[1]CONCEDE *Consortium for Central European Dictionary Encoding*, http://www.itri.brighton.ac.uk/projects/concede/

[2]MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages*, http://nl.ijs.si/ME/

computed as the ratio between the number of the occurrences of the respective lemma and the number of the occurrences of the most frequent lemma of that POS. Therefore the normalized ranking of a lemma is a real number less or equal to 1 (it is 1 only for the most frequent lemma). For each occurrence of an inflected form of a given lemma, the respective lemma was credited with one more occurrence. The frequency range figures were computed for each part of speech, so that we could select for each part of speech high, medium and low frequency words of the respective category. The proper names and abbreviations were discarded from the selection process (usually, they are not proper items for explanatory dictionaries). 562 lexical entries from the Bulgarian Explanatory Dictionary (Andreychin et al. 1994), covering the word list produced according to the above-mentioned procedure, were selected. The number is slightly greater than 500 because the dictionary contained multiple entries for homographs. It includes some reference entries as well. These 562 lexical entries contain information for 591 lemmas, because some of the entries contain more than one lemma (for instance, masculine and feminine forms for some nouns).

The chosen entries are divided in the following POS:

Noun	200	33.84%
Verb	130	21.99%
Adjective	74	12.52%
Adverb	68	11.51%
Total (open)	472	79.86%
Numeral	9	1.52%
Pronoun	31	5.24%
Conjunction	24	4.06%
Preposition	21	3.55%
Particle	26	4.40%
Interjection	8	1.35%
Total (closed)	119	20.13%
Total	591	99.99%

Second, the next 5500 lemmas were selected upon the following principal breakdown of lemmas to parts of speech (agreed by the CONCEDE consortium): open POS (nouns, verbs, adjectives, adverbs) — no more than 90%, closed POS (numerals, pronouns, conjunctions, prepositions, particles and interjections) — minimum 10% out of the whole set of lemmas chosen. **Encoding scheme:** the Bulgarian and Polish (like CONCEDE project languages) use different character sets (Cyrillic for Bulgarian and Latin with some special characters for Polish). That's why the Bulgarian-Polish LDB uses 8-bit encoding defined in the Unicode 8.

4. Web-based Application for the presentation of the Bulgarian-Polish online dictionary

The next task required a design and development of the LDB of Bulgarian-Polish online dictionary. We used the CONCEDE LDB (Erjavec, Evans, Ide, Kilgarriff 2000) as a model, and extended the monolingual CONCEDE LDB to a bilingual one (for more on the Bulgarian-Polish LDB see Dimitrova, Panova, Dutsova 2009).

The technologies used for the implementation of the web-application for the presentation of the Bulgarian-Polish online dictionary are Apache, MySQL, PHP and JavaScript. We employ free technologies originally designed for developing dynamic web pages with a lot of functionalities. With the help of HTML and CSS we design and create the both "administrative" and "end-user" modules. The "super-administrator", who has the rights to manage the dictionary' content, manages the "administrative" module. The software tool offers a user-friendly interface for adding, editing, deleting and searching words. The access to this module is restricted and only people who have authorization can access it.

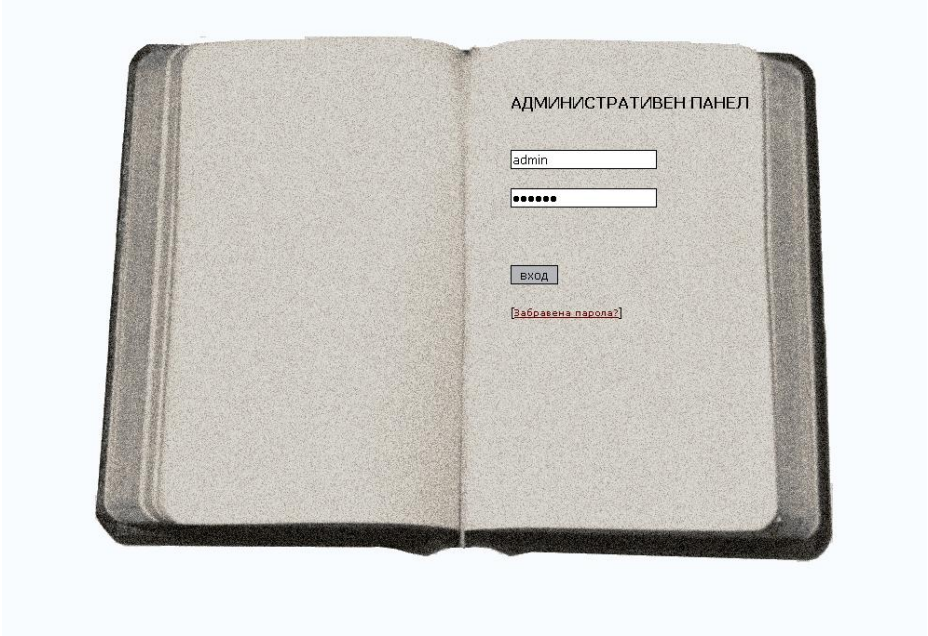

Figure 1. Administrative module — login in the web-application

Let us illustrate how the administrative module works. We consider an example entry whose headword appears in the list of selected Bulgarian headwords, and show the steps which one "administrator" should follow to upload a new entry to the LDB supporting Bulgarian-Polish online dictionary.

For this purpose we choose the Bulgarian verb **разбирам** /understand/. We choose a verb, because the verbs are the richest POS with specific characteristics. In

Bulgarian, a very well developed system exists for expression of the "tense"-category: there are forms to express 9 different verb tenses. The verb also supports expression of the following grammatical categories: person, number, voice, aspect, tense and mode. Depending on particularities of their lexical meaning, Bulgarian verbs are classified as either transitive (allow a direct subject — the action is transferred from the subject to another object), or intransitive (the action is not transferred to an object). In traditional printed dictionaries not all specifications are encoded and presented by the respective classifiers. To represent the Bulgarian verbs more adequately in the online dictionary we have included some additional information about conjugations' type. We introduce "transitive" and "intransitive" syntactic classifiers (in this case transitivity refers to the usage of nouns as direct objects following the verbal form). Semantic information related to the aspect forms of the verbs was also introduced, namely a new semantic classifier to mark the meaning of the verbal form with values state and event.

The Bulgarian verb **разбирам** /*understand*/ could be found in the Slawski dictionary (Sławski 1987):

разби'ра|м, -ш *vi.state, transitive*; rozumieć *transitive*; ∼**м от не'що** znam się na czymś; ∼**м бъ'лгарски** rozumiem po bułgarsku; ∼ **се** rozumie się, ma się rozumieć; staje się jasne, zrozumiałe; ∼**м се** *aux.* rozumieć się, porozumiewać się, godzić się

We illustrate next how the above example will be uploaded in the Bulgarian-Polish online dictionary.

After the username and user's password have been entered and verified (**Figure 1**), the user is redirected to the administrative module. The administrative module contains one of several different sections: one for new word entry, one for searches of Bulgarian or Polish words, one for translation setting, etc.

The user must choose from a combo box what he/she wants to enter: a noun, verb, adjective or any other POS (pronouns, conjunctions, adverbs), in this case — a verb. The fields displayed then are only the ones necessary for adding a verb. All fields needed for entry of other POS would be hidden for the user.

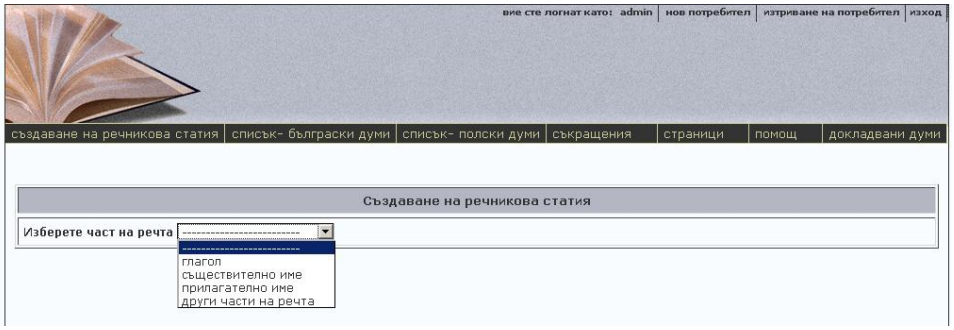

Figure 2. Administrative module — choosing the type of the new word to be added

At the second step of verb upload, the headword-box will be filled in with "**разбирам**", the grammatical characteristics of the verb "**разбирам**" — its conjugation in the 2nd person, singular; its conjugation type: I, II or III can be chosen from a drop-down list; whether the verb is "transitive" or "intransitive"; as well as the "perfect aspect" (vp) or "imperfect aspect" (vi), and the semantic feature of the verb expressing a "state" or "event" can also be chosen from other drop-down lists. Some explanations about the determination of the conjugation type of any verb and the definition of transitivity/intransitivity of verbs can be found in the "Help"-section of the administrative module.

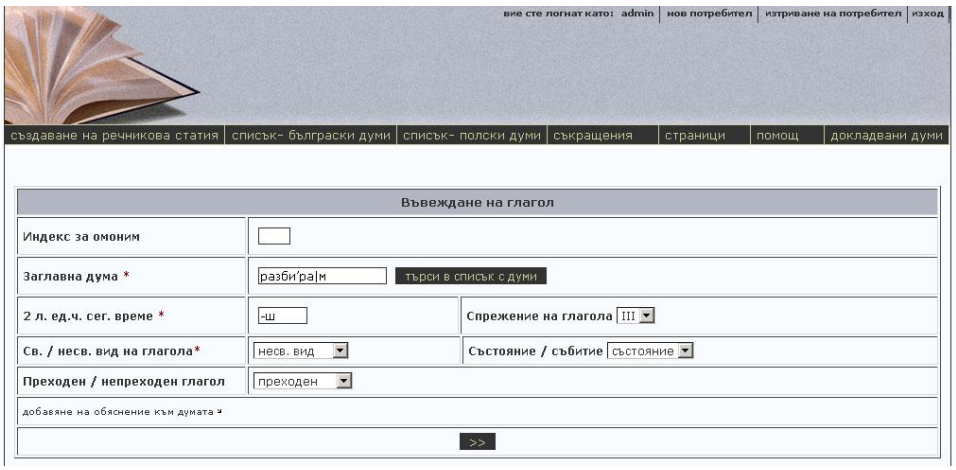

Figure 3. Administrative module — adding the grammatical characteristics of the verb **разбирам**

When all the information is filled in, the administrator would press the Next ">>" button. At the next step, a new form is displayed, where the administrator will enter information about a ***specific use*** of the verb, such as its use as a "medical term", "botanical term", etc., and/or any ***stylistic meanings*** (archaic, folklore, etc.). At this step, if it is necessary, references to another word can be created.

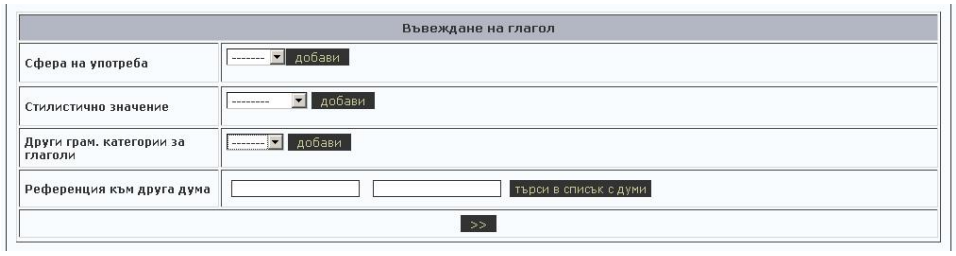

Figure 4. Administrative module — addition of stylistic meanings and the creation of reference

At the third step, the administrator will fill in the text fields the corresponding Polish translations (meanings) of the Bulgarian word. Using the "add" button, the administrator can add multiple Polish translations. A "drop-down" list which can be used to give detailed information about the Polish verbs usage is also included. For some Polish verbs (but not all) one can have the transitive/intransitive classifier as well.

In our example, there is only one meaning (with one classifier) in Polish.

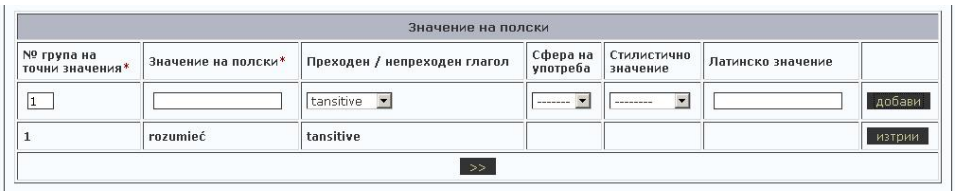

Figure 5. Administrative module — adding a Polish translation (meanings)

For each POS there is a common part that ensures the possibility to add an unspecified number of derivations, phrases and examples for each headword. At the forth step, the administrator must add examples, derivations and phrases for the current verb.

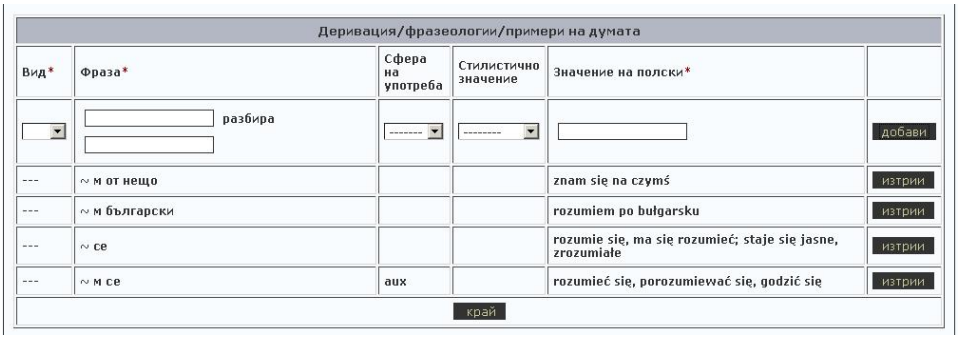

Figure 6. Administrative module — adding examples, derivations and phrases for the verb *разбирам*

This is the last step of the verb entry. When the administrator presses the "Finish" button, the word is added in the LDB, and it will be possible to search for and display it in the "user-end" module. Within the administrative module, there are the foreseen possibilities to edit and delete an already existing dictionary entry; also to add, delete, and update all kinds of characteristics, abbreviations and their explanations. Through the "Help" menu the user can add more topics to enrich the user manual or to read the already existing ones

5. Functionality of the "end-user" module

The "end-user" module is "bilingual": the user can choose the input language (Bulgarian or Polish). There is the possibility to search for a translation in both directions: from Bulgarian to Polish and from Polish to Bulgarian. The translation from Bulgarian to Polish will display the whole information that exists in the LDB of the dictionary for the searched word. The translation from Polish to Bulgarian will be composed only using the main senses of the Bulgarian headwords. The "end-user" module provides a **Contact form** where the casual user can report words currently missing in the dictionary or to warn about errors or gaps in the dictionary entry.

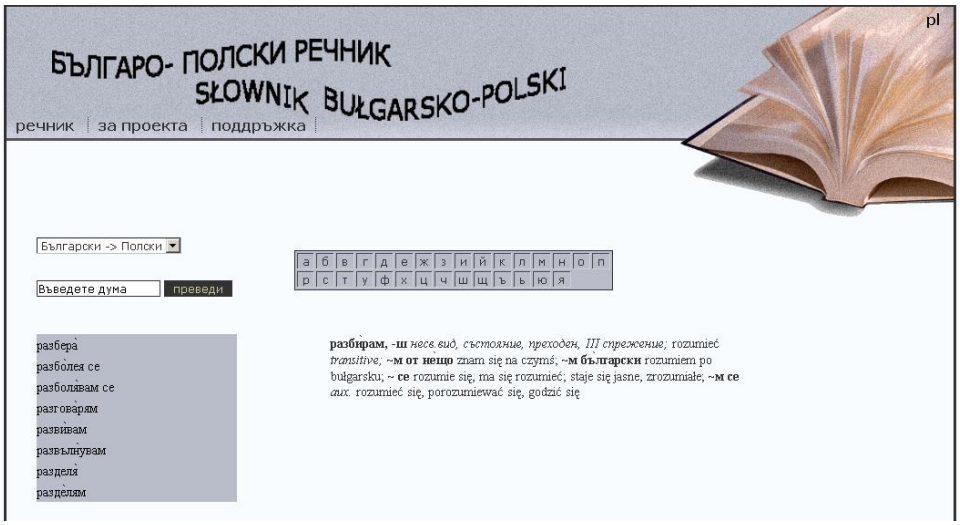

Figure 7. "End"-user module — translation of the Bulgarian word *разбирам*

Figure 8. End-user module — translation of the Polish word *rozumieć*

In a blue block on the left-hand side, words that are alphabetically closest to one the searched and currently available in the LDB of the dictionary will get displayed. The "end-user" can choose the input language (Bulgarian or Polish), and a virtual Bulgarian or Polish keyboard is displayed according to their choice. This way, the user can choose special Bulgarian or Polish characters if they are not supported by different keyboards.

When the user chooses Bulgarian to Polish translation, the whole information saved in LDB is displayed. When translating from Polish to Bulgarian, only the Bulgarian headwords are visualized with their possible grammatical characteristics.

6. Conclusion

In this paper, we present the recent developments of the Bulgarian-Polish online dictionary. The dictionary is still at an experimental stage and is intended for research purposes only, but can be useful in the daily life, for educational and translation purposes.

Some suggestions for improvement the dictionary follow:

Extending the dictionary is a feasible task. The established Bulgarian-Polish parallel corpus, which contains more than 3 million words, can provide a good basis for a lexical dictionary. The main difficulty in the implementation of bilingual electronic dictionaries, where the transfer takes place in both directions, is that in any language the lexical forms have more than one meaning and do not overlap in two-way translation. The first Bulgarian-Polish dictionary has the potential to develop, grow and become a widely available and useful tool.

References

Andreychin et al. (1994): Andreychin, L., Georgiev, L., Ilchev, St., Kostov, N., Lekov, I., Stoikov, St., Todorov, Tsv. Bulgarian Explanatory Dictionary. 4th revised edition. Dimitar G. Popov editor. Sofia, Nauka i Izkuvstvo Publishing House, 1093 pages. (In Bulgarian)

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

Dimitrova, L., Dutsova, R., Panova, R. (2011). Survey on Current State of Bulgarian-Polish Online Dictionary. In: *Proceedings of the International Workshop "Language Technologies for Digital Humanities and Cultural Heritage"* within International Conference RANLP'2011, 16 September 2011, Hissar, Bulgaria. 43–50.

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

Dimitrova, L., Panova, R., Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. In: *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, 36–47.

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

Erjavec, T., Evans, R., Ide, N., Kilgarriff, 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

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.

Sławski F., (1987). *Podręczny słownik Bułgarsko-Polski z suplementem.* 2nd edition, Warszawa, Polska.

Tufis, D., Rotariu, G., Barbu, A.-M. (1999). Data sampling, lemma selection and a core explanatory dictionary of Romanian. In: *Proceedings of COMPLEX'99,* Pecs, Hungary, 219–228.