

LUDMILA DIMITROVA^{1,A} & RALITSA DUTSOVA^{1,B}

¹Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia

^Aludmila@cc.bas.bg ; ^Br.dutsova@yahoo.com

WEB-APPLICATION FOR THE PRESENTATION OF BILINGUAL CORPORA (FOCUSING ON BULGARIAN AS ONE OF THE TWO PAIRED LANGUAGES)

Abstract

This paper briefly presents a web-application for the presentation of bilingual aligned corpora focusing on Bulgarian as one the two paired languages. The focus is given to the description of the software tools and user interface. The software is developed in IMI-BAS and will be hosted on a server there. Some examples of the usage of the web-application for the presentation of a Bulgarian-Polish aligned corpus are included.

Keywords: parallel corpus, aligned corpus, concordance, linguistic annotation, lemmatization, POS-tagging, web-interface, web-application.

1. Introduction

The software tool *Web-application for the presentation of bilingual aligned corpora with Bulgarian* focuses on pairs of languages with Bulgarian being one of the two. The texts in the ongoing version of the corpora are automatically aligned at the sentence level. The whole corpus is oriented towards emphasizing the applicability of the digital bilingual data for computerized natural language processing, but also as a source of human readable information.

2. Format of the Texts

The bilingual aligned corpora using Bulgarian as one of the paired languages, prepared at the Mathematical Linguistics Department of the IMI-BAS under the supervision of L. Dimitrova, will serve as input files for the software tool *Web-application for the presentation of bilingual aligned corpora with Bulgarian*.

2.1. Alignment of a Corpus

For a parallel corpus to be useful, it must be treated with a special program for "alignment". An aligned corpus is a parallel corpus containing relations between

corresponding chunks of text of multiple languages (Dimitrova, Garabík, 2011), (Keliš, 2009), (Moore, 2002), (Rosen, 2005), (Varga et al., 2005). The alignment is a process of relating pairs of words, phrases, sentences or paragraphs in the texts in different languages which are translation equivalent. Commonly, parallel corpora are aligned at the sentence level, because the alignment aims to produce a set of corresponding sentences (original and its translation(s)). (One of the most well-known examples of parallel text alignment is inscribed on the famous Rosetta stone.) The result of the alignment of two parallel texts is a merged document, usually called bi-text, composed of both source- and target-language versions of a given text that retains the original sentence order. The alignment tools are software tools that generate bi-texts: they align automatically the original and translated versions of the same text. The tools generally match these two texts sentence by sentence. Our decision here is in favour of pair-wise alignments, rather than of a table-like alignment (applicable to the paragraph-level alignment of texts). In addition, "alignment" is a form of annotation carried over parallel corpora to facilitate the construction and evaluation of translation models stored in memory and used in support of computer-assisted translation. Although many parallel corpora are manually "aligned", automatic "alignment" forms the core of parallel corpora processing and tool development for "alignment" with a high degree of accuracy.

We used language-independent freely-available software tools to align bilingual corpora in which Bulgarian was one of the two languages: the MT2007 Memory Translation computer aided tool (TextAlign), and the Bitext Aligner/Converter (Bitext2tmx aligner). TextAlign is a software package that segments and aligns corresponding translated sentences, contained in two rich text format files. Bitext2tmx aligner is a Java application. It works on any Java supported operating system (e.g. Windows, Linux, Mac OS X, Solaris), and is released under the GNU General Public License. Bitext2tmx aligner is a software tool that segments and aligns corresponding translated sentences, contained in two plain text files. These software packages have applications in computer-assisted translation. Both tools align bilingual texts without bilingual dictionaries, but human editing is obligatory. The resulting aligned texts are similar.

The following example presents an excerpt from the aligned at the sentence level (1561 bilingual links in total) Bulgarian and Polish texts of *The Little Prince*, French writer Antoine de Saint-Exupéry's most famous novella (using TextAlign):

```
<tu tuid="0000000022">
  <tuv lang="Bulgarian">
    <seg>Така че трябваше да избира друг занаят и се научих да управ-
лявам самолети.</seg>
  </tuv>
  <tuv lang="Polish">
    <seg>Musiałem wybrać sobie inny zawód: zostałem pilotem.</seg>
  </tuv>
</tu>
<tu tuid="0000000023">
  <tuv lang="Bulgarian">
```

```

    <seg>Летял съм по малко навсякъде по света. И наистина, географията много ми помогна.</seg>
  </tuv>
  <tuv lang="Polish">
    <seg>Latałem po całym świecie i muszę przyznać, że znajomość geografii bardzo mi się przydała.</seg>
  </tuv>
</tu>
<tu tuid="0000000024">
  <tuv lang="Bulgarian">
    <seg>От пръв поглед можех да различа Китай от Аризона.</seg>
  </tuv>
  <tuv lang="Polish">
    <seg>Potrafiłem jednym rzutem oka odróżnić Chiny od Arizony.</seg>
  </tuv>
</tu>

```

The next table shows an excerpt of the aligned at the sentence level (6699 bilingual links in total) Bulgarian-English Orwell's 1984 texts, so-called Orwell corpus (Dimitrova et al., 2005), (using Vanilla Aligner):

<Obg.1.1.10.1> Уинстън рязко се обърна.	<Oen.1.1.11.1> Winston turned round abruptly.
<Obg.1.1.10.2>Беше надянал маската на спокоен оптимизъм, която бе препоръчителна за пред телекрана.	<Oen.1.1.11.2>He had set his features into the expression of quiet optimism which it was advisable to wear when facing the telescreen.
<Obg.1.1.10.3>Прекося стаята и влезе в кухненския бокс.	<Oen.1.1.11.3>He crossed the room into the tiny kitchen.

2.2. Annotation

Corpus annotation is the process of adding linguistic or structural information to a text corpus. The annotations make the corpora more useful for linguistic research. One common type of annotation is the addition of labels or tags that indicate the word class for the words in the text. This is the so-called part-of-speech tagging (POS tagging): information about each word's part of speech (verb, noun, adjective, adverb, etc.) in the form of labels — tags — is added to the words of the corpus. Another example of linguistic annotation is indicating the lemma — the base form of each word (Dimitrova, Koseska-Toszewa, Derzhanski, & Roszko, 2009). Apart from linguistic annotations, there are other types of annotation, for example, structural annotations, which correspond to different structural levels of a corpus or text. Written texts contain a number of different structural forms or divisions. Novels have a complex hierarchy and are divided into parts and chapters, newspapers are divided into sections, reference works — into articles, etc. The most

common division in this hierarchy is the paragraph. The structural annotation allows the texts in the two languages (Bulgarian and Lang2) to be aligned at the corresponding level in order to produce aligned bilingual corpora.

Some texts in the ongoing version of the Bulgarian-Lang2 corpora (Bulgarian-English, Bulgarian-Lithuanian, Bulgarian-Polish, Bulgarian-Russian, Bulgarian-Slovak) are annotated at the paragraph level, others are aligned and therefore annotated at the segment level (usually the sentence level). The standard markers `<p>` and `</p>` for a paragraph's boundaries, `<seg>` and `</seg>` for a segment's boundaries, are employed. The `<p>` level alignment allows the drawing of a broader context in the languages. In other words, there is the opportunity — thanks to the broader context — to more precisely study the meanings of word-forms in each language.

3. Design Decisions and Web Interface

3.1. Web-application

The technologies used for the implementation of the *Web-application for the presentation of the bilingual aligned corpora* are Apache, MySQL, PHP and JavaScript. We use free technologies originally designed for developing dynamic web pages with many functionalities. The software tool offers a user-friendly interface for adding to, editing, deleting from and searching the database. The following web-based application is experimental, and the structure of the text fields is not permanently determined yet. Changes are possible during the implementation process.

3.2. Query interface

The *Web-application for the presentation of bilingual aligned corpora* consists of two parts — an administrative and an end-user part. The administrative part has a very simple interface and offers the possibility for the user to add to, edit, delete from and search the database of the corpus. After the administrator has logged in to the system, he/she is redirected to a page where a sentence or a word contained in the sentence can be searched by identification. After a search has been performed, the user has the ability to edit the database if corrections are needed or delete the listed records.

The application provides a simple web form where the user can insert a new pair of aligned texts:

```
<tu tuid="0000000023">
  <tuv lang="Bulgarian">
    <seg>Летял съм по малко навсякъде по света. И наистина, географията много ми помогна.</seg>
  </tuv>
  <tuv lang="Polish">
    <seg>Latałem po całym świecie i muszę przyznać, że znajomość geografii bardzo mi się przydała.</seg>
  </tuv>
```

вие сте логнат като: admin нов потребител изтриване на потребител изход	
Търсене	
Добавяне на нов запис	
Добавяне на записи	
ID *	<input type="text" value="0000000023"/>
БГ текст	<input type="text" value="Летял съм по малко навсякъде по света. И наистина, географията много ми помогна."/>
ПЛ текст	<input type="text" value="Latałem po całym świecie i muszę przyznać, że znajomość geografii bardzo mi się przydała."/>
<input type="button" value="запази"/>	

Figure 1. Insertion of a new pair of aligned texts.

The web-based end-user interface is bilingual. Only a search-by-word capability is provided to the end user. The user can choose the input language (Bulgarian or Lang2 — Polish, in this case). A virtual keyboard is implemented to help the end user insert search criteria more easily in case he/she does not have the Bulgarian or Lang2 alphabet installed on the computer used.

The search is performed according to the primary language selected by the user. All pairs of aligned text where the searched word has been found are listed in a table. In order to show the word in a better context, together with the target pair we display the previous and next pair as well. If the search results exceed more than 15 records, paging is provided.

3.3. Relational Database, Supporting Web-application

The base of the Web-application is the relational database (RDB) of the Bulgarian-Lang2 (Polish in this case) corpora. The relational model is supported by tables containing core information of the corpora entries and the links established between them.

The usage of RDB for storage of the corpora entries has several advantages: maintenance of the integrity of data, ensuring data security and independence, quick and efficient search and data retrieval, upload and updates. We paid special attention to building the database that supports the web presentation of bilingual corpora in order to address the following computational complexities.

Searching a large text can be a costly operation, one that takes up a long time to run. The database structure was therefore designed in a way to provide easy and fast search capabilities for the end-users of the bilingual web corpora.

When a user inserts a new record in the database through the administrative module, a backend text parser program takes the input text and simplifies it to its separate constituent words. The different words are then saved in different fields in an index table of the database, and for each word a link is kept to another table

where the full text of the aligned pair is saved. This parsing is done for both texts — the one in Lang2 as well as the Bulgarian text.

In this way, we achieve a good search performance and only a small delay while inserting new records in the database. The delay is not so sensible and the administrator will not pay a big attention to it, because he has the possibility to add the new aligned pairs only one by one.

Description of the Tables of the Database corpus_db — the Base of the Web bilingual corpora presentation on the example of the Bulgarian-Polish aligned corpus:

sentence_bg_pl — contains the information about the aligned Bulgarian-Polish (i.g. Lang2) pair.

word_bg — after the Bulgarian text has been parsed, each Bulgarian word from the aligned pair is saved in this field of database.

word_pl — after the Polish text has been parsed each Polish word from the aligned pair is saved in this field of database.

word_sentences_bg — for each record in the table **word_bg**, a link is kept to the table **sentence_bg_pl**. The combination **id_word_bg**, **id_sentence_bg_pl** is unique.

word_sentences_pl — for each record in the table **word_pl**, a link is kept to the table **sentence_bg_pl**. The combination **id_word_pl**, **id_sentence_bg_pl** is unique.

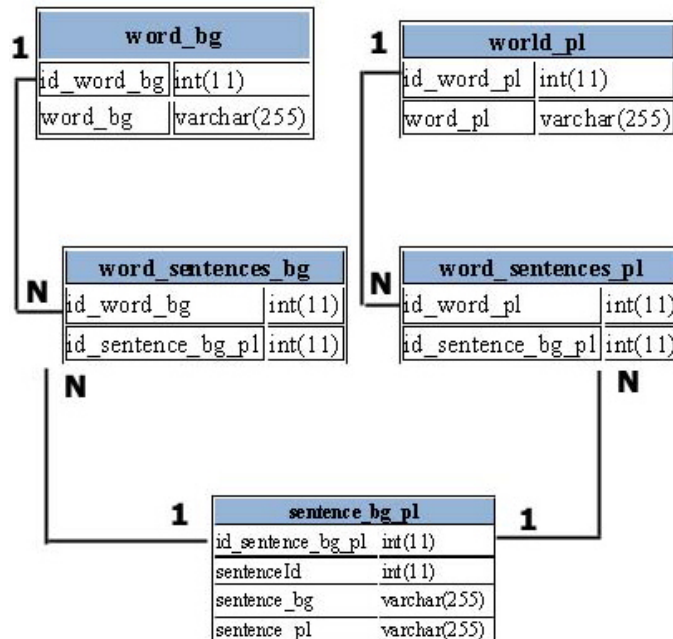


Figure 2. Structure of the Relational Database.

3.4. Concordances

The digital corpora are the main base of knowledge in corpus linguistics, and the aligned corpora are the best resource for the development of different kinds of digital dictionaries and other special type of lists, namely concordances. A concordance is a list of occurrences of a given (specified) word or phrase used in given large text, a book or a corpus, together with their immediate context.

The text retrieval and concordancing programs as a computer enhanced tools, provide the linguists an opportunity of studying vast amounts of text in a short time. 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. The citations thus obtained can be sorted to reveal recurring clusters of words (Dimitrova, Garabik, 2011). The analysis of these recurring patterns highlights the behavior of actual language in context. That is why, the concordances have many applications in contrastive studies: they are used for comparison of different uses of the same word (in a different context) and for creating indexes and lists of words; in a keyword analysis and analysis of the frequency of words; to locate and analyze phrases and idioms in a given text; to find the translation of the essential elements of text, such as terms (in multilingual texts).

The Figure 3 shows a concordance with the Bulgarian noun „света” /world/ (from A. de Saint-Exupéry’s *The Little Prince*):

ЗАЯВКА

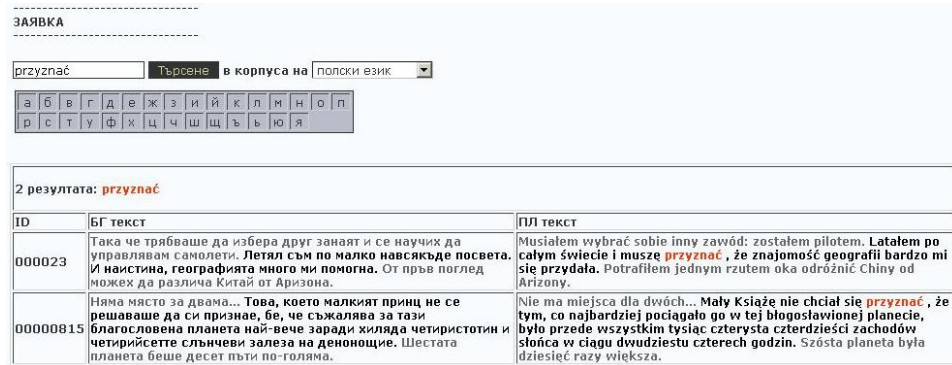
света Търсене в корпуса на

а б в г д е ж з и й к л м н о п
р с т у ф х ц ч ш щ ъ њ ю я

7 резултата		
ID	БГ текст	ПЛ текст
000023	Така че трябваше да избира друг занаят и се научих да управлявам самолети. Летял съм по малко навсякъде по света. И наистина, географията много ми помогна. От пръв поглед можех да различа Китай от Аризона.	Musiałem wybrać sobie inny zawód: zostałem pilotem. Latałem po całym świecie i muszę przyznać, że znajomość geografii bardzo mi się przydała. Potrafiłem jednym rzutem oka odróżnić Chiny od Arizony.
000043	Първата вечер заспах върху пясъка, на хиляди мили от всякаво населено място. Вях откъснат от света повече от корабкрушенец върху сал сред океана. Сигурно си представяте моята изненада, когато сутринта не събуди странно гласче.	Pierwszego wieczoru zasnąłem na piasku, o tysiąc mil od terenów zamieszkałych. Byłem bardziej osamotniony, niż rozbitek na tratwie pośrodku oceanu. Toteż proszę sobie wyobrazić moje zdziwienie, gdy o świecie obudził mnie czyjś głosik.
000339	Не е по-сериозно и по-важно от сметките на един дебел червен господин? И ако аз познавам едно-единствено цвете на света, което не съществува никъде освен на моята планета, а една малка овца може някоя сутрин да го унищожи с едно движение ей така, без да си дава сметка какво прави, това не е важно! Той се изчерви, после продължи:	Czy to nie jest ważniejsze niż rachunki grubego, czerwonego pana? Jeżeli ja znam jedyny kwiat, który nigdzie poza moją planetą nie istnieje, i jeżeli mały baranek może go któregoś ranka zniszczyć za jednym zamachem, nie zdając sobie sprawy z tego, co czyni, czyż nie ma to żadnego znaczenia? Poczzerwieniał. Po chwili mówił dalej:
001070	Но ако не опитомиш, ние ще изпитваме необходимост един от друг. За мен ти ще бъдеш единствен на света. За теб аз ще бъда единствена на света.	Lecz jeżeli mnie oswoisz, będziemy się nawzajem potrzebować. Będziesz dla mnie jedyny na świecie. I ja będę dla ciebie jedyny na świecie.
001071	За мен ти ще бъдеш единствен на света. За теб аз ще бъда единствена на света. - Започвам да разбирам - каза малкият принц.	Będziesz dla mnie jedyny na świecie. I ja będę dla ciebie jedyny na świecie. - Zaczynam rozumieć -- powiedział Mały Książę.
001146	- Печеля - отговори лисицата - заради цвета на житото. И добави: - Иди да видиш отново розите. Ще разбереш, че твоята е единствена на света. После се върни да се сбогуваме и ще ти подаря една тайна.	- Zyskałem coś ze względu na kolor zboża -- powiedział lis, a później dorzucił: -- Idź jeszcze raz zobaczyć róże. Zrozumiesz wtedy, że twoja róża jest jedyna na świecie. Gdy przyjdiesz pożegnać się ze mną, zrobię ci prezent z pewnej tajemnicy.
001153	Беше само лисица, подобна на сто хиляди други. Но я направих моя приятелка и сега е единствена на света. И розите много се смутиха.	Był zwykłym lisem, podobnym do stu tysięcy innych lisów. Lecz zrobiłem go swoim przyjacielem i teraz jest dla mnie jedyny na świecie. Róże bardzo się zawstydzily.

Figure 3. Concordances with the Bulgarian noun „света” /world/ in the corpus.

The Figure 4 shows a concordance with the Polish verb *przyznać* /admit/ (from A. de Saint-Exupéry's *The Little Prince*):



ID	БГ текст	ПЛ текст
000023	Така че трябваше да избира друг занаят и се научих да управлявам самолети. Летял съм по малко навсякъде посвета. И наистина, географията много ми помогна. От пръв поглед можех да различа Китай от Аризона.	Musiałem wybrać sobie inny zawód: zostałem pilotem. Latałem po całym świecie i muszę przyznać, że znajomość geografii bardzo mi się przydała. Potrafiłem jednym rzutem oka odróżnić Chiny od Arizony.
00000815	Няма място за двама... Това, което малкият принц не се решаваше да си признае, бе, че съжалява за тази благословена планета най-вече заради хиляда четиристотин и четирийсетте слънчеви залеза на денонощие. Шестата планета беше десет пъти по-голяма.	Nie ma miejsca dla dwóch... Maly Książę nie chciał się przyznać, że tym, co najbardziej pościągalo go w tej błogosławionej planecie, było przede wszystkim tysiąc czterysta czterdzieści zachodów słońca w ciągu dwudziestu czterech godzin. Szósta planeta była dziesięć razy większa.

Figure 4. Concordances with the Polish verb *przyznać* /admit/ in the corpus.

4. Applications of Web-presented Bilingual Corpora

The aligned parallel corpora are useful for many NLP applications: in systems for machine-aided human translation, in systems for machine translation for the training of software tools, for training of translators. They are prerequisite for contrastive studies or other linguistics research, and can also be used for retrieval of linguistic information, for producing concordances, for developing bi- and multilingual lexical databases and different kinds of digital dictionaries, etc. ((Dimitrova, Koseska, 2009), (Dimitrova et al., 2010), (Garabik, Dimitrova, Koseska, 2011)).

4.1. Training of Translators (Humans or Programming Tools)

The main application area of the corpora is the translation. The parallel and aligned corpora are successfully used as a translation database and language materials for the training of translators, humans or programming tools for machine translation, as well as in education — for language learning in schools and universities. Instructors, students and professional translators can use bilingual corpora as a complementary resource in educational process.

Bilingual corpora can also be used for training of software packages for automatic disambiguation of morphosyntactic annotation.

The advantage of processing a text corpus is to obtain context specific information about syntactic structures and usage of words in a given language. In the case of parallel corpora, one can obtain context-specific correlations between these languages, which are usually much less ambiguous than general collections. Resulting data from these corpus analysis processes can be used to develop context-specific tools for translation and to standardize the usage of structures and word sets for future multilingual document production.

This approach is more correct as there is no "word"-to-"word" comparison, but a comparison of word-forms in a broader context, which allows a better identification of the word's meaning.

4.2. Development of Multilingual Lexical Databases and Digital Dictionaries

The parallel and aligned corpora are the best resource for the development of bi- and multilingual lexical databases and different kinds of digital dictionaries.

Multilingual parallel corpora represent a good base of data for bilingual dictionaries creation. There are many research projects for automatic extraction of bilingual lexical knowledge from parallel corpora in the field of information retrieval from large scale text corpora. Thus parallel corpora are successfully used for automatic lexicon extraction. There one could find and extract many examples of the usage of the words from the corpus in a wide context.

4.3. Applications in Contrastive Studies

Every language in the bilingual corpus — Bulgarian, English, Lithuanian, Polish, and Russian — exhibit some specific features, occurring repeatedly in several categories. At first, different orthography traditions — the corpora are dataset of written languages and the orthography forms an inseparable part of language analysis. A significant feature is the analytic character of Bulgarian and English. In the process of evolution of Bulgarian from a synthetic, inflectional language, to an analytic, flectional language, case forms were replaced with combinations of different prepositions with a common case form. Bulgarian has lost most of the traditional Slavic case system, exhibits several linguistic innovations in comparison to the other Slavic languages (a rich system of verbal forms, a definite article), and has a grammatical structure closer to English or the Neo-Latin languages than other Slavic languages. One of the most important grammatical characteristics of the Bulgarian language which sets it apart from the rest of the Slavic languages is the existence of a definite article. Bulgarian possesses high number of verbal forms and a strongly developed category of verbal aspect ((Dimitrova, Koseska, 2012), (Dimitrova, Koseska, Roszko, D., & Roszko, R., 2009a, 2009b, 2010, 2011)).

The web-presented bilingual aligned corpora are oriented both to human and machine users and are available for a wide area of applications: corpora and frequency lists derived from them are useful for language teaching. Recently, the aligned corpora serve as a basis for development of new applications in multilingual digital libraries.

Conclusion

There are many projects for automatic extraction of bilingual lexical knowledge from parallel corpora in the field of information retrieval from large scale text corpora. Aligned corpora are the most effective means for the creation of bi- and multilingual dictionaries and contrastive grammars. One has to remember that parallel corpora comprise direct material for the evaluation of translations and their analysis will bring out the improvement of the quality of both traditional, human translation, and machine translation. Besides, texts extracted from parallel or aligned corpora prove the necessity of evaluating translations: it is common that in translation words get omitted or word meanings get changed. That is why online free-use aligned texts are a useful education resource.

Future development and implementation. In a future implementation of the web application, texts in the bilingual corpora will be lemmatized or/and POS tagged. The bilingual aligned corpora with Bulgarian will be freely available for research and education on the web with an appropriate tri- or bilingual interface in Bulgarian, Lang2, and English (when Lang2 is not English). The Bulgarian-Polish aligned corpus will be soon represented via Internet by the *Web-application for presentation of bilingual corpora with Bulgarian* with a wide spectrum of features for practical applications.

References

- Dimitrova et al. (2010). L. Dimitrova, V. Koseska, R. Garabík, T. Erjavec, L. Iomdin, V. Shyrovkov. *Conceptual Scheme for a Research Infrastructure Supporting Resources in Slavic Lexicography*. Sofia, Demetra Ltd. Publisher, pp. 131. ISBN 978-954-8986-33-5.
- Dimitrova, L., Garabík, R. (2011). Bulgarian-Slovak Parallel Corpus. In: *Proceedings of the Sixth International Conference NLP, Multilinguality SLOVKO'2011, 20–22 October 2011*, Modra, Slovakia, p. 44–55. ISBN 978-80-263-0049-6
- Dimitrova, L., Koseska, V. (2009). Bulgarian-Polish Corpus. *Cognitive Studies / Études cognitives*, 9, SOW, Warsaw, p. 133–141. ISSN 2080-7147. DOI: 10.11649/cs.2009.010
- Dimitrova, L., Koseska, V. (2012). Bulgarian-Polish Parallel Digital Corpus and Quantification of Time. *Cognitive Studies / Études cognitives*, 12, SOW, Warsaw, 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: V. Shyrovkov, L. Dimitrova (Eds.), *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, V., Roszko, D., & Roszko, R. (2009a). Bulgarian-Polish-Lithuanian Corpus — Problems of Development and Annotation. In: T. Erjavec (Ed. 2009), *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, 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 web-page of the Association for Computational Linguistics (ACL), <http://www.aclweb.org/anthology/W/W09/W09-4001>).
- Dimitrova, L., Koseska, 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, SOW, Warsaw, p. 217–240. ISSN 2080-7147. DOI: 10.11649/cs.2010.013
- Dimitrova, L., Koseska, 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 et al. (2005). Dimitrova, L., R. Pavlov, K. Simov & L. Sinapova. Bulgarian MTE Corpus — Structure and Content. *Cybernetics and Information Technologies*, 5(1), p. 67–73.
- Garabík, R., Dimitrova, L., Koseska, V. (2011). Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). *Cognitive Studies / Études cognitives*, 11, SOW, Warsaw, p. 227–239. ISSN 2080-7147. DOI: 10.11649/cs.2011.014
- Ide Nancy (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In: First International Conference on Language Resources and Evaluation, LREC'98, p. 463–470, Granada, ELRA. <http://www.cs.vassar.edu/CES/>
- Kelih, E. (2009). Slawisches parallel-textkorpuz: Projektvorstellung von „kak zakaljalas' stal' (kzs)“: In: Kelih, E., Levickij, V. & Altmann, G. (Eds.), *Metody analizu teksta/Methods of Text Analysis*, p. 106–124, Černivci. ČNU.
- Moore, R. (2002). Fast and accurate sentence alignment of bilingual corpora. In: *Machine Translation: From Research to Real Users*, p. 135–144.
- Rosen, A. (2005). In search of the best method for sentence alignment in parallel texts. In: Garabík, R. (Ed.), *Computer Treatment of Slavic and East European Languages: Third International Seminar, Bratislava 10–12 November 2005*, p. 174–185, Bratislava.
- Varga et al. (2005). Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. Parallel corpora for medium density languages. In: *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, p. 590–596.

Bitext2tmx: <http://bitext2tmx.sourceforge.net>

TextAlign: <http://mt2007-cat.ru/index.html>

