Radovan Garabík[1], Ludmila Dimitrova[2], Violetta Koseska-Toszewa[3]
[1]garabik@kassiopeia.juls.savba.sk, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
[2]ludmila@cc.bas.bg, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
[3]amaz@inetia.pl, Institute of Slavic Studies, Polish Academy of Sciences

Web presentation of Bilingual Corpora (Slovak–Bulgarian and Bulgarian–Polish)

[1]Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia,
[2]Institute of Mathematics and Informatics, Sofia, Bulgaria,
[3]Institute of Slavic Studies, Warsaw, Poland

Abstract
In this paper we focus on the web-presentation of bilingual corpora in three Slavic languages and their possible applications. Slovak-Bulgarian and Bulgarian-Polish corpora are collected and developed as results of the collaboration in the frameworks of two joint research projects between Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, from one side, and from the other side: Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences and Institute of Slavic Studies, Polish Academy of Sciences, coordinate by authors of this paper.
Keywords: Bulgarian, Polish, Slovak, digital language resources, parallel and aligned corpora, web presentation

1. Introduction

The main objectives of creating Slovak-Bulgarian and Bulgarian-Polish corpora are connected to the collaboration in the framework of the joint research projects between Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, from one side, and from the other side: Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences and Institute of Slavic Studies – Polish Academy of Sciences, coordinate by authors of this paper. The first project "Electronic Corpora – Contrastive Study with Focus on Design of Bulgarian-Slovak Digital Language Resources" focuses on design, collection and development of the first Bulgarian-Slovak and, respectively, Slovak-Bulgarian parallel corpora. The second project "Semantics and Contrastive Linguistics with a Focus on a Bilingual Electronic Dictionary" focuses on the design and development of bilingual electronic dictionary based on the first Bulgarian-Polish corpus. These digital bilingual resources will be widely applicable to the contrastive studies of Slavic languages. The aligned at the paragraph and sentence level parallel corpora will be valuable resources for machine translation research.

2. Preliminary works

For Bulgarian language the first parallel corpus has been produced as a part of the MULTEXT-East[1] corpus consisting of George Orwell's "1984" in English and its translations in six CEE languages. The alignment of sentences from the English original with the Bulgarian text is also a part of the MULTEXT-East corpus (Dimitrova et al. 1998).

For Slovak language, the first parallel corpus was a Russian-Slovak parallel corpus, created as a collaboration project between Ľ. Štúr Institute of Linguistics and Faculty of Philology, St. Petersburg State University (Garabík, Zakharov 2006), and the second corpus was French-Slovak corpus (Vasilišinová, Garabík 2007). Both corpora feature automatic lemmatization and morphosyntactic annotation, automatic alignment using the Hunalign software (Varga et

[1] The EU COP Project 106 MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages,* http://nl.ijs.si/ME/

al. 2005) and a web interface providing CQL queries over the Manatee-based backend. The interface has been upgraded and extended for the use in subsequent Czech-Slovak and Slovak-Bulgarian parallel corpora.

For Polish language the first parallel Bulgarian-Polish corpus has been developed as a collaborative work under the joint research project "Semantics and Contrastive Linguistics with a Focus on a Bilingual Electronic Dictionary" (Dimitrova, Koseska 2009).

3. Common differences between the languages

At first we will stress on the differences resulting from orthography tradition, since we are primarily dealing with the written language, where the orthography forms an inseparable part of language analysis. Bulgarian uses the Cyrillic alphabet, Polish and Slovak – Latin with additional special symbols and diacritical marks.

There are **some features specific for three languages**, which occur repeatedly in several categories, and which we mention here at the beginning. Some specific morphosyntactic characteristics of these three Slavic languages was described in more details in (Dimitrova et al. 2009a), and (Dimitrova et al. 2009b).

A significant feature is the analytic character of Bulgarian, and the synthetic character of Slovak and Polish. 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. Bulgarian 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 Polish or Slovak. 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. The definite article is a morphological indicator of the grammatical category determination (definiteness). The definite article is not a particle, nor is it a simple suffix, but a meaningful compound part of the word. It is a word-forming morpheme, which is placed at the end of words in order to express definiteness, familiarity, acquaintance. In Bulgarian, nouns, adjectives, numerals, and full-forms of the possessive pronouns and participles can acquire an article. Polish and Slovak lacks the definiteness attribute altogether.

The above mentioned specific characteristics reflect to the different level of annotations and should be taken in mind in preparation of the annotated parallel and aligned multilingual resources.

4. Description of Corpora and Current Development

4.1. Slovak-Bulgarian corpus – Parallel and Aligned

The Bulgarian–Slovak corpus (currently under development) comprises two corpora: parallel and aligned.

The Bulgarian–Slovak/Slovak– Bulgarian parallel corpus contains more than 1 200 000 words, mostly fiction, novels, and short stories. The main part of parallel corpus contains texts in other languages translated into both Bulgarian and Slovak.

The aligned corpus, 376 200 words, contains parallel texts, aligned at the paragraph level and at the sentence level. The set of aligned texts includes Bulgarian novels: Dimitar Dimov's *Doomed Souls* and Pavel Vezhinov's *The Barrier* and their Slovak translations, the novel of

Slovak writer Klára Jarunková *The silent wolf's brother* and its Bulgarian translation, Bulgarian and Slovak translation of Jaroslav Hašek's *The Good Soldier Švejk*. The language-independent freely-available program Hunalign that aligns parallel texts at the sentence level is used. The program foresees the use of a corresponding bilingual dictionary to ensure a higher accuracy of the alignment. The result of aligning Bulgarian and Slovak translation of The Good Soldier Švejk without a dictionary is fully satisfactory:

„ Aby to išlo do počtu , do tucta , lepšie sa to ráta a na tucty je to vždy lacnejšie , “ odpovedal Švejk .

- За по - лесно , като са дузина , по - лесно се броят , пък и на дузини всичко е по - евтино - отговори Швейк .

4.2. Bulgarian-Polish corpus – Parallel, Aligned, and Comparable

The corpus is built with the main purpose to ensure the selection of the entries for the first experimental electronic Bulgarian-Polish dictionary. The texts were collected concurrently and do not have a connection with national monolingual or other corpora.

The Bulgarian–Polish corpus consists of two corpora: a parallel and a comparable. All collected texts in the corpus are texts published in and distributed over the Internet and were downloaded from existing digital libraries.

A detailed description of the corpus is provided for clarification to the user. The description includes: language, author, title, words in the text, and if available, year of creation, publication place, year and publishing house, translator, year of translation, source and original format of the text, etc.

A part of the parallel texts is annotated at paragraph level. A small part of the corpus is currently aligned at sentence level and forms so-called aligned Bulgarian-Polish corpus. This approach is more correct – we are not comparing "word" with "word", we compare word-forms in a broader context, which allows us to obtain the word's meaning.

The Bulgarian–Polish parallel corpus includes two parallel sub-corpora a core and a translated:

1) A core Bulgarian–Polish parallel corpus consists of original texts in Polish (literary works by Polish writers and their translation in Bulgarian) and original texts in Bulgarian (short stories by Bulgarian writers and their translation in Polish).

2) A translated Bulgarian–Polish parallel corpus consists of texts in Bulgarian and in Polish of EC brochures, EU and EU-Parliament documents, published online; Bulgarian and Polish translations of third language literary works (mainly English).

The Bulgarian–Polish comparable corpus includes texts in Bulgarian and Polish with the text sizes being comparable across the two languages: excerpts from textual documents, shown online, excerpts from several original fiction, novels or short stories. Some of the Bulgarian texts are annotated at “paragraph” (level <p>) and “sentence” (level <s>) levels, according to CES (Ide et al. 2000).

The aligned corpus includes texts of Polish novels: Stanisław Lem's Solaris and Return from the Stars, Ryszard Kapuściński's The Shadow of the Sun and Another Day of Life, and Stefan Żeromski's Ashes and their Bulgarian translations. The two language independent freely available programs Memory Translation 2007, a computer aided tool (TextAlign: http://mt2007-cat.ru/index.html) and Bitext Aligner/Converter (bitext2tmx aligner: http://bitext2tmx.sourceforge.net/) that aligns parallel texts at the sentence level is used. Bitext Aligner/Converter, a Java application, is a program to align and segment corresponding

translated sentences, contained in two plain text files, and generate a translation memory (TMX format) from them for use in computer-aided/assisted translation applications. An example from aligned at sentence level Lem's novel Return from the Stars follows:

```
<tu tuid="0000000039">
    <tuv xml:lang="Polish">
        <seg>Zostawałem bezustannie w tyle za wszystkim, co się działo, i ciągłe usiłowanie zrozumienia byle rozmowy, sytuacji zmieniało to napięcie w uczucie paskudnie podobne do rozpaczy.</seg>
    </tuv>
    <tuv xml:lang="Bulgarian">
        <seg>Все изоставах назад от всичко, което се случваше край мен, и непрекъснатото старание да разбера кой да е разговор и ситуация превръщаше това напрежение в нещо отвратително, подобно на отчаяние.</seg>
    </tuv>
```

5. Bilingual corpus on the Web

The source texts are kept in original format, each source (typically one book) in its own directory, with two subdirectories, one for each language. Each source text has a conversion script that converts the source format into plain UTF-8 text file. If necessary, the file is then manually edited to remove parts not in correspondence between the languages – typically the foreword or other editorial remarks, so that we get two files with corresponding beginnings and ends. The Slovak text is then lemmatized and morphosyntactically tagged, the files are aligned and using the information about rung alignment, vertical files with structural tags numbering the sentences and their aligned counterparts are created. The vertical files are indexed for the Manatee corpus manager, each of the languages providing one corpus.
When querying the corpus, for each concordance the sentence(s) from the opposite language corpus are identified and displayed in tabular format (Fig.1). All the standard Manatee CQL features are available – regular expression on the character level, regular expressions on the token level, searching for different attributes (lemma and morphosyntactic tag in Slovak), searching within structures (sentences).

```
<doc lang="sk" origlang="sk"
               biblio="Odsúdené duše,
                           Slovenský spisovateľ,
                           Bratislava 1976,
                           preklad Milan Topoľský">
<s id="1" link="1">
Odsúdené       odsúdený       Gtfp1x 12
duše           duša           Ssfp1  04
</s>
<s id="2" link="2">
Dimitär        dimitär        %      01
Dimov          dimov          %      24
</s>
<s id="3" link="3">
Prvá           prvý           Nafs1  02
časť           časť           Ssfs1  03
Koniec         koniec         Ssis1  04
jedného        jeden          Nfns2  04
dobrodružstva  dobrodružstvo  Ssns2  01
</s>
```

Fig. 1. Example of Slovak vertical file, wordforms with lemmas, morphosyntactic tags and number of disambiguation possibilities, sentences with links into Bulgarian text.

The question of reusability has long been a key issue of digital language resources. It is well known that the development of such resources is a lengthy process; however, this development is all too often repeated again and again, because ready-made resources are not available in a usable format or their distribution is hindered due to propetry rights, diverse and poorly documented encodings, unwillingness of the corpus owners/producers to distribute them further. For this reason we used TEI to circumvent some of these restrictions, however limitations remain, namely, those are due to copyright restriction exercised on the corpus annotation, i.e., on the corpus as a whole, was well as on the component texts.

6. Applications of Slovak–Bulgarian and Bulgarian–Polish Corpora

The web-presented language resources are oriented both to human and machine users and are available for a wide area of applications. The parallel bilingual corpora, aligned at the paragraph or the sentence level, annotated in accordance with international standards provide samples of the words meaning and usage in different context when digital dictionary are developed. In addition, these corpora are useful as a language material for bilingual lexical and terminological databases and on-line dictionaries development (Dimitrova et al. 2009c, Šimkova et al. 2009). The corpora uploaded on the web could be used successfully in education. It can also be used as a translation database and language learning materials for training of translators – human and programming tools.

The bilingual corpora have special applications in contrastive studies on Slavic languages. As already mentioned in (Dimitrova, Koseska 2009) a parallel corpus of whatever languages cannot be a sum of separate monolingual corpora of the given languages. Simultaneous accumulation of equivalent texts must be observed across the various languages. Monolingual corpora contain material illustrating the diversity and different levels (synchronous and diachronous) of the development of a language system. Parallel corpora have to contain language material that must be equivalent albeit translated, that is synchronous. Parallel corpora cannot take into account the diachronous level in the language development, and that level requires a different approach at the annotation of material and does not help the development of multilingual dictionaries nor cannot be of any help to machine translation.
Concerning the systems for digital corpus annotation we must note that the annotation of a bilingual parallel corpus requires the usage either of the same system of annotation tags, reflecting the particularities of both languages, or two comparable (via a one-to-one correspondence) sets of annotation tags for each language.

An additional problem in the development of bilingual parallel corpora is caused by the proportion of translated literature in the chosen languages. For example, Polish literature is more often translated into Bulgarian than otherwise.

Our hope is to develop and improve the system for digital corpus annotation based on our experience. From the viewpoint of parallel corpora application to the discussion of language problems, we stress that the languages are not randomly chosen: the three languages belong to the Slavic group. Bulgarian belongs to the South-Slavic group, Polish and Slovak to the West-Slavic, but Slovak is a West Slavic language with many characteristics of the South Slavic group.

Slovak is characterised by the change of the Protoslavic phonetic sequence **-ort-**, **-olt-** into **-rat-**, **-lat-**, which is characteristic of the South-Slavic subgroup. Instead of the sequence **-rot-**, **lot-** in Polish, Slovak contains expressions with the sequence **-rat-**, **-lat-** as in Bulgarian and the other South-Slavic languages: compare for example the Slovak nouns **rakyta**, **laket'** and the Bulgarian nouns **ракита**, **лакът** with the Polish **rokita**, **łokieć.**

So far the search tool realized searches in Slovak texts for a lemma and in Bulgarian or Polish texts – for a wordform. Let's use our corpora to demonstrate an application in contrastive studies. The following examples from the Slovak-Bulgarian parallel aligned corpus (Dimitar Dimov's *Doomed Souls* and the Slovak translation, see the Fig. 2) and the Bulgarian-Polish parallel corpus (see the excerpt form Kapuściński's novel, Fig. 3) demonstrate the usage of the Bulgarian noun „**лакът**":

28427	Chová sa celkom prirodzene . Veľa nechýbalo a boli by sa pohádali , ale mních dovečeral a oprel sa **lakťami** o stôl . Bolo zrejmé , že chce odísť , no zo slušnosti sa rozhodol , že sa s nimi chvíľu pozhovára .	Човекът се държи съвсем нормално . Бяха почти готови да се скарат , когато монахът свърши вечерята си и се облегна с **лакът** върху масата . Стана явно , че искаше да си иде , но от вежливост бе решил да разговаря няколко минути с тях .
89970	„ Nervy už mám nanič , " povedala takmer nahlas . Keď zhášala cigaretu , zhodila **lakťom** baterku , a ako utekala , zachytila sa bedrom o nohu mŕtveho . Najlepšie bude , keď pôjde von .	„ Нервите ми са разстроени " — каза тя почти гласно . Гасейки цигарата си , бе съборила с **лакът** електрическото фенерче от масата и като тичаше навън , бе закачила с бедрото си крака на умрелия . Най - добре беше да излезе от палатката .
103367	Pod jej zdrapmi mu bolo vidno telo , belasé od bitky . Mal skrvavenú tvár i **lakeť** , červenovlasú hlavu hrdo vztýčenú a jeho pery sa pohrdlivo usmievali na revúci dav . Mlčal . Jeho druhovia , jednoduchí , nevzdelaní ľudia oplácali zástupu nadávky a s typickou španielskou nezmieriteľnosťou sa vyhrážali tým , ktorí vrieskali najväčšmi .	Под дрипите на ризата се виждаше тялото му , посиняло от бой . Лицето и единият му **лакът** бяха окървавени ; червенокосата му глава стоеше гордо вирната нагоре и устните му се усмихваха презрително към ревящата тълпа . Той мълчеше , докато другарите му , хора прости и неуки , ругаеха множеството , зъбеха се и с чисто испанска непримиримост се заканваха на тия , които викаха най - много .

Fig. 2. Web search interface for Slovak-Bulgarian corpus

Pl-Bg	Zbiegłem po schodach na dół, gdzie w recepcji wsparty **łokciami** o stosy niepotrzebnych papierów i sterte pieniedzy bez wartosci drzemał Felix, j ego blada twarz le.ała na dłoni nieruchoma, bez wyrazu.	Изтичах по стълбите долу, където в рецепцията, подпрян с **лакти** на купчини непотребни документи и пачки пари без никаква стойност, дремеше Фелиш - бледото му лице бе опряно на ръката му, неподвижно, безизразно.

Fig. 3. The excerpt from Ryszard Kapuściński's *Another Day of Life*

Another transformation of the Proto-Slavic phonetic sequence **-tort-**, **-tolt-** in these languages is an important phonetic criterion for the classification of the Slavic languages. In the West-Slavic Polish language this sequence is replaced by the sequence **-trot-, -trót-,** in East-Slavic we have **-torot-**, **-tolot-** respectively, and in the South-Slavic **-trat-**, **-tlat-**. In Slovak and Czech these sequence this sequence is also transformed into **-trat-**, **-tlat-**, like in the South-Slavic languages, see for instance the usage of **млад** //young// and **здрав** //healthy// in the following context (part of Slovak-Bulgarian corpus).

3728	- Možno ani nevieš , že absolvoval medicínu , filozofiu a teológiu . Bol ešte **mládenec** a už mal v hierarchii rádu vysokúhodnosť . Padre Pedro , gróf Sandoval , ho považoval za svojho zástupcu a budúceho generála rádu .	Може би не знаеш : свърши медицина , философия и теология … Още съвсем **млад** стигна висока степен в йерархията на ордена . Отец Педро , графът на Сандовал , го сочеше като свой заместник и бъдещ генерал на ордена …
29582	Chlap , ktorý sa tak surovo dožadoval rovnosti , sa objavil vo dverách . Bol **mladý** , čiernovlasý , zavalitý a mal dobrácke hnedé oči . ♦ Ich výraz vôbec nezodpovedal jeho srditým výkrikom , ktorými akoby štiepal budovu . Mal čistú pracovnú kombinézu , a ak by sa malo usudzovať z toho , ako vykrikoval , bol trocha podpitý .	Човекът , който така свирепо настояваше за равенство , се показа на вратата . Той бе **млад** , чернокос , с грамадно телосложение и добродушни кафяви очи , чийто израз никак не отговаряше на сърдитите викове , с които цепеше посадата . Носеше чистичък работнически комбинезон и ако се съдеше по виковете му , бе малко пийнал .
47470	Vládne orgány ich zatkli , ale na druhý deň zaútočil na väzenie dav a mládencov ubili palicami na smrť . V Cadíze vypískalo obecenstvo **mladého** , opatrného , neskúseného torera . „ Podliak ! . . .	Властта ги арестува , но на другия ден тълпата нападна затвора и уби младежите с тояги . В Кадикс публиката освирка предпазливостта на един **млад** , неопитен тореро . „ Подлец ! …

Fig. 4. Some appearances of the Bulgarian adjectives „**млад**" //young// in Slovak-Bulgarian corpus.

15526	Luíz sa zachvel . Taká dávka bola pre **zdravého** človeka smrteľná , ale pre ňu - celkom normálna . Odmeral požadované množstvo , a keď jej podal striekačku s tampónom vaty , namočeným v liehu , odvrátil hlavu .	Луис потрепера . Дозата спокойно можеше да умъртви **здрав** човек , но за нея бе нормална . Той отмери исканото количество и като й подаде спринцовката с памука , натопен в спирт , извърна глава .

Fig. 5. An appearance of the Bulgarian adjectives „**здрав**" //healthy// in Slovak-Bulgarian corpus.

It is important to note that the Bulgarian language participating in both corpora can serve as a intermediate language and help in the comparison of the three languages. A well-known fact is that the Bulgarian language is typologically different from Polish and Slovak. It has an analytical structure, while Polish and Slovak are synthetic. The comparison of Polish and Slovak with Bulgarian gives an opportunity for scholars to study the modern tendency to analytism in two synthetic Slavic languages. The tendency to analytism appear in the mixing of some case endings, whereby these cases have disappared in Bulgarian already during the 9-10 century.

The comparison of the two aligned corpora illustrates well how the meaning of the Bulgarian definite article is conveyed in Polish and Slovak. Due to the analytism in Bulgarian, the category „definiteness – indefiniteness" is not only expressed analytically but also lexically – like in Polish and Slovak. Different types of pronouns play a special role in expressing the meanings "uniqueness" (definiteness) and "universality and existentiality" (i.e. indefiniteness) of the object. Bulgarian language material will focus the researcher attention to the temporal and modal meanings expressed at the level of the semantic structure of the phrase in these three languages. The comparison of the language material of the two parallel corpora reveals the common inadequacy of some Bulgarian-to-Slovak and Bulgarian-to-Polish translations. For examples, the traditional Slovak and Polish grammars do not pay attention to the modality, expressed by grammatical means in Bulgarian and called „imperceptiveness". The translators overlook this modal meaning following the principle „this grammatical form is missing in the language, hence the semantic phenomenon is missing and has no translation"! Let us compare the following sentence (Fig. 6):

Pl-Bg	Znaleźli pewnego starego Anglika, który **postanowił** (***past perfect tense expressed by the form of praeterium***) stąd wyjechać przy najbliższej okazji i chce sprzedać łódź motorową w dobrym stanie.	Бяха намерили един стар англичанин, който **бил решил** (***form of encreasy imperceptiveness expressing imperceptive modality***) да се маха оттук при първия възможен случай и искал да продаде моторна лодка в добро състояние.

Fig. 6. The excerpt from *The Shadow of the Sun* by Ryszard Kapuściński

Being an analytical language, Bulgarian has preserved a great number of verbal forms for past and future tense. Of special interest are the meanings of aorist form of imperfective verbs and the meanings of imperfect form of imperfective verbs. The both forms differ in their temporal meaning although their verbal aspect is the same. These differences in meaning are described in detail in (Koseska 2006), (Koseska & Mazurkiewicz 2010). Traditional grammar do not cope with the description of the temporal meanings of these two forms. It is no surprise that the translators do not reflect upon their temporal differences and translate the text from Bulgarian to Polish and Slovak with the same form for different meanings. This is a serious error which we discover easily thanks to the parallel corpora. Let us compare the following translations presented in Fig. 7 and 8.

87944	Nato sa pobrala zobudiť Robinsona . Keď v rannom šere mních sadol na bicykel a navždy opustil tábor , Fanny **pozerala** za ním , až kým sa jeho postava nezmenila na bodku a nezmizla na belavom páse hradskej . Nechcela ho zadržať kvôli Herediovi ?	И тръгна да събуди Робинзон. Когато в здрача на зората монахът яхна колелото и напусна завинаги лагера, Фани дълго **гледа** след него, докато фигурата му се превърна в точка и най - сетне изчезна в белезникавата лента на шосето . Не искаше ли да го задържи заради Ередиа ?

Fig. 7. The usage of **глед`а** (aorist form of imperfective aspect of verb in Bulgarian)

Pl-Bg	Gdy się na pewien przeciąg czasu rozjaśniło, Rafał głośniej począł wykrzykiwać wiersze Horacego, zbliżył się cicho do szyb i długo, z wytężoną uwagą w dal **patrzał** (***form of praeterit of imperfective aspect of the verb in Polish***).	Когато за кратко време се развидели, Рафал почна да скандира стиховете на Хораций, приближи се тихо до прозореца и дълго, с напрегнато внимание **глед`а** (***form of aorist of imperfective aspect of the verb in Bulgarian***) в далечината.

Fig. 8. The example from the Bulgarian-Polish corpus, Stefan Żeromski's „Ashes", shows the usage of the same Bulgarian verb

The next examples (Fig. 9 and 10) show the usage of **гледаше** (imperfect form of verbs of imperfective aspect in Bulgarian):

91448	- Mŕtvolu nikto neukradne . Dievčina poslúchla , vošla do stanu a očami rozšírenými od strachu **pozrela** na Fanny . - Čo na mňa tak hľadíš ?	Никой няма да открадне мъртвеца … Девойката се подчини и влезе в палатката , като **гледаше** Фани с широко разтворени от уплаха очи . — Не ме гледай така ! — меко каза Фани , като я погали по бузата . —

Fig. 9. An excerpt from the Slovak-Bulgarian corpus

Pl-Bg	Przez chwilę stał tam u wejścia na gołoborze i, przejęty głuchą trwogą, **patrzał** (***praeterit form of imperfective aspect of the verb in Polish***) w to miejsce, gdzie wiedźmy, strzygi, błędnice zlatują się o północku i gdzie się ukazuje sam Zły.	Известно време стоя там, на границата, където почваше голо бърдо, и завладян от смътна тревога, **гледаше** (***imperfect form of verbs of imperfective aspect in Bulgarian***) към мястото, където вещици, вампири и бродници се събират в полунощна доба и се явява сам Злият.

Fig. 10. An example from the Bulgarian-Polish corpus, Stefan Żeromski's „Ashes"

Of course, there are also good translations, where such modal nuances are conveyed by the respective language means, for example in Stefan Żeromski's *Ashes*:

Pl-Bg	Radzono tedy, co czynić. Jedni utrzymywali, że wypadnie objechać aż na Koprzywnicę; inni byli zdania, żeby się cofnąć w górę rzeki, gdzie **jakoby** przed laty **był** jakiś lekki mostek.	Едни смятаха, че ще трябва да се заобиколи чак през Копшивнице; други бяха на мнение, че ще трябва да се върнат нагоре по реката, където като че ли преди години **имало** (***imperceptiveness)*** някакво леко мостче.

The translation errors in the above examples can be avoided by means of a semantic analysis in the modern comparative grammars such as the Bulgarian-Polish contrastive grammar, where the description goes in the direction "meaning-form" and is made through a semantic intermediary language, so that Polish and Bulgarian are equivalently-described languages, see (Koseska 2006), (Koseska 2009), (Koseska & Mazurkiewicz, 2010). In this grammar it does not matter whether Polish or Bulgarian stays in first place. This is a grammar for both languages.

The distinction between morphological, syntactic and lexical level in traditional grammars carries over to the annotation of language corpora, leading scholars to formal description of natural languages. Normal and machine translations, however, are highly dependent on annotations showing the meanings of forms in the languages compared. This has to be a requirement for annotation sets especially in parallel corpora. Only such type of annotation tags will help avoid the above-mentioned translation errors and orient the scholars towards semantics and language comparison.

6. Conclusion

The paper describes the presentation on the web the one million words Slovak–Bulgarian and the three millions words Bulgarian–Polish parallel aligned corpora. Parallel aligned corpora are a language resource for contrastive, translation and terminology studies, for development of machine translation and other multilingual technologies, like tools for development of

lexical databases and digital dictionaries. Special attention has been given to enabling further distribution of the corpora by encoding them in a standard format.

Further work will involve enriching the annotation of the corpora that will make them more representatives, as regards composition and size.

References

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, pp. 315-319.

Dimitrova, L., Koseska-Toszewa, V. (2008). *Some problems in multilingual digital dictionaries*. Études Cognitives. Vol. 8, SOW, Warsaw, 2009, pages 237–254.

Dimitrova, L., Koseska, V. (2009). *Bulgarian-Polish Corpus*. Cognitive Studies/Études Cognitives. Vol. 9, SOW, Warsaw, 2009, pages 133–141.

Dimitrova et al. (2009a): Dimitrova, L., Garabík, R., Majchráková, D. Comparing Bulgarian and Slovak Multext-East morphology tagset. In: Shyrokov, Volodymyr & Dimitrova, Ludmila (Editors, 2009). Organisation and Development of Digital Lexical Resources. *Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009.* Dovira Publ. House, Kiev, 2009, pages 38-46.

Dimitrova et al. (2009b): Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. Bulgarian-Polish-Lithuanian Corpus – Problems of Development and Annotation. In: Erjavec (Editor, 2009), Research Infrastructure for Digital Lexicography. *Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana.* Informacijska družba Publ. House, Ljubljana, 2009, pages 72-86.

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

Garabík et al. (2004): Garabík, R. Gianitsová, L., Horák, A., Šimková, M., Šmotlák, M. Slovak National Corpus. In: *Proceedings of the conference TSD 2004. Brno, Czech Republic*. Springer-Verlag.

Garabík, R., Špirudová, J. (2009). Design of a New Slovak-Czech Lexical Database. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, pages 71–76.

Garabík, R., Zakharov, V. P. (2006). Parallel Russian-Slovak Corpus. In: *Proceedings of the International Conference on Corpus Linguistics*. St. Petersburg State University Publishing House. St. Petersburg, 2006, pages 81 – 87. (In Russian)

Ide et al. (2000): Ide, N., Bonhomme, P., and Romary, L. XCES: An XMLbased Encoding Standard for Linguistic Corpora. *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: ELRA, 2000. pages 825-830.

Koseska–Toszewa, V. (2006). Semanticzna kategoria czasu, Gramatyka konfrontatywna bułgarsko – polska, t. 7 , Warszawa, SOW, 210 pages. (In Polish)

Koseska–Toszewa, V. (2009). Many-volume Contrastive Grammar of Bulgarian and Polish. In: Shyrokov, Volodymyr & Dimitrova, Ludmila (Editors, 2009). Organisation and Development of Digital Lexical Resources. *Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009.* Dovira Publ. House, Kiev, 2009, pages 87-97.

Koseska–Toszewa, V., Mazurkiewicz, A. (2010). *Time Flow and Tenses*. SOW, Warsaw, 2010, 223 pages.

Šimkova et al. (2009): Šimková, M., Garabík, R., Dimitrova, L. Design of a multilingual terminology database prototype. In: Koseska, Violetta, Dimitrova, Ludmila, Roszko, Roman (Editors, 2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open Workshop, Warsaw, Poland, 31 May–2 June 2009,* SOW, Warsaw, 2009, 123–127.

Stieber, Z. (1974). Świat językowy Słowiań. Warswzawa 1974, PWN, pages 145 - 146. (In Polish)

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In: *Proceedings of the Recent Advances in Natural Language Processing, 21-23 September 2005, Borovets, Bulgaria (RANLP'2005).* INCOMA Ltd. Shoumen, Bulgaria, 2005, pages 590–596.

Vasilišinová, D., Garabík, R. (2007). Parallel French-Slovak Corpus. In: Computer Treatment of Slavic and East European Languages. *Proceedings of the conference Slovko 2007*. Eds. J. Levická, R. Garabík. Brno: Tribun 2007.