MAKSIM DUŠKIN¹, JOANNA SATOŁA-STASŃKOWIAK²

¹Institute of Slavic Studies, Polish Academy of Sciences, Warsaw
(duszkin.m@gmail.com)

²Institute of Slavic Studies, Polish Academy of Sciences, Warsaw
(joannasat@op.pl)

THE BULGARIAN-POLISH-RUSSIAN PARALLEL CORPUS

Abstract

The Semantics Laboratory Team of Institute of Slavic Studies of Polish Academy of Sciences is planning to begin work on the creation of a Bulgarian-Polish-Russian parallel corpus. The three selected languages are representatives of the main groups of Slavic languages: Bulgarian represents the southern group of Slavic languages, Polish — the western group of Slavic languages, Russian — the eastern group of Slavic languages. Our project will be the first parallel corpus of these three languages. The planned corpus will be based on material, dating from one period (the 20th century) and will have a synchronous nature. The project will not constitute the sum of the separate corpora of selected languages.

One of the problems with creating multilingual parallel corpora are different proportions of translated texts between the selected languages, for example, Polish literature is often translated into Bulgarian, but not vice versa.

Bulgarian, Russian and Polish differ typologically — Bulgarian is an analytic language, Polish and Russian are synthetic. The parallel corpus should have compatible annotation, while taking into account the characteristic features of the selected languages.

We hope that the Bulgarian-Polish-Russian parallel corpus will serve as a source of linguistic material of contrastive language studies and may prove to be a big help for linguists, translators, terminologists and students of linguistics. The results of our work will be available on the Internet.

Keywords: parallel corpora, the Bulgarian-Polish-Russian parallel corpus

1. About the Bulgarian-Polish-Russian parallel corpus.

It is not necessary today to convince anyone of the need for corpus research or the positive effects of this research used in linguistics (but not only!). Studies concerning corpus linguistics, methods of action as well as simple tools¹ to use corpora can be found, e.g. on the Internet. The latest thematic academic conferences such

¹ For example: *www.korpusy.net*

as, e.g. “Slavicorp” also reflect a great interest in corpus linguistics. It is the first international conference organised by Polish Studies Department of Warsaw University, the Polish Language Foundation and the Polish Language National Corpus and we also participated in it².

Every parallel corpus contains a basis of texts and their translations. Such a corpus allows to establish the most appropriate way of translating a given linguistic unit into another language. By means of the corpus the translator can, for example, take into consideration various contexts in which translation variants of a specific unit occur and choose the most suitable one.

A semantics team of the Institute of Slavic Studies of the Polish Academy of Sciences directed by Violetta Koseska, encouraged by work in the international project Mondilex³, is approaching the end of works on bilingual electronic Bulgarian-Polish dictionary. The project concentrated on maintaining the multilingualism and multiculturalism of Europe by the creation of a general scheme of research infrastructure supporting lexicographical research on Slavic languages and the development of new or already existing digital stores of lexical items according to the world standards. Therefore the team wants to take up a new challenge, an attempt to create **the Bulgarian-Polish-Russian parallel corpus**.

The prospects of creating, as we hope, valuable tools in science, which would constitute a source of linguistic knowledge as well as practical knowledge essential for a contemporary translator and researcher, are exceptionally encouraging. The effects of this work can direct the choice of the most suitable methods of contemporary applied linguistics.

We are familiar with parallel or comparable corpora which are based on Slavic linguistic material (e.g. IPI PAN⁴, NKJP⁵, Bul -Pol Korpus⁶). The authors of

² M. Duškin, J. Satoła-Staškowiak, About the Bulgarian-Polish-Russian parallel corpus project, Slavicorp 2010

^{3,4} The partners in the project are research organisations from six European countries whose six national languages belong to the Slavic group: Bulgaria (Institute of Mathematics and Informatics, Bulgarian Academy of Sciences), Poland (Department of Semantics of the Institute of Slavic Studies, Polish Academy of Sciences), Russia (Institute for Information Transmission Problems, Russian Academy of Sciences), Slovakia (L. Štúr Institute of Linguistics, Slovak Academy of Sciences), Slovenia (Jožef Stefan Institute) and Ukraine (Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine). All partners are national centres for high-quality research in lexicography and digital resources. (see www.mondilex.org)

⁴ The IPI PAN Corpus is a large (currently over 250 million segments), morphosyntactically annotated, publicly available corpus of Polish, developed by the *Linguistic Engineering Group* at the *Institute of Computer Science, Polish Academy of Sciences* (ICS PAS), mainly within projects funded by the *State Committee for Scientific Research*, as well as within statutory research carried out at ICS PAS. (<http://korpus.pl/>)

⁵ The National Corpus of Polish is a shared initiative of four institutions: *Institute of Computer Science* at the Polish Academy of Sciences (coordinator), *Institute of Polish Language* at the Polish Academy of Sciences, *Polish Scientific Publishers PWN*, and the Department of Computational and Corpus Linguistics at the *University of Łódź*. It has been registered as a research-development project of the *Ministry of Science and Higher Education*. (<http://nkjp.pl/>)

⁶ The Parallel Corpus mentioned in the text additionally consists of two equivalent Bulgarian-Polish subcorpora: in the first one the annotation proceeds on the level of ‘sentence’ in literature (fiction), and in the second case on the level of paragraph — specialized vocabulary (text documents of the European Commission and the European Union, including the third language — English). (Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. Bulgarian-Polish-Lithuanian

the last from the above mentioned- the experimental Bulgarian-Polish Comparable and Parallel Corpus are: V. Koseska⁷ i L. Dimitrova⁸. At present the Corpus has already 3 000 000 word forms isolated from the Bulgarian and Polish literature as well as from text documents of the European Commission and the European Union, containing specialized vocabulary, including the third language — English.

Our project of **the Bulgarian-Polish-Russian Parallel Corpus** would be the first parallel corpus of these specific languages.

Formulating the aim — the parallel corpus, we mean only these texts in one language which have been translated into the two remaining languages of the corpus. We are interested in juxtaposing the language of the original with the texts that are its translations. The planned corpus will be based on the material from one period (the 20th century) and will have a synchronous character. The project will not constitute the sum of the separate corpora of selected languages.

We hope to observe many interesting linguistic phenomena resulting from the translation of the examined texts and their juxtaposition.

The Bulgarian-Polish-Russian Parallel Corpus will serve as a source of linguistic material of contrastive language studies and may prove to be a big help for linguists, translators, terminologists and students of linguistics. The results of our work will be available on the Internet.

The three deliberately selected languages are representatives of the main groups of Slavic languages: Bulgarian represents the southern group of Slavic languages, Polish — the western group of Slavic languages, Russian — the eastern group of Slavic languages. The people working in our project (V. Koseska, M. Duškin, J. Satoła-Staškowiak,) are natural carriers of one of the languages of the corpus. We hope that it will additionally help in work on specific texts entered into the corpus and annotated for its needs.

The problem that arises with creating multilingual parallel corpora are different proportions of translated texts between the selected languages, for example Polish literature is often translated into Bulgarian but not vice versa.

Corpus–Current Development. In: Proceedings of the International Workshop “Multilingual resources, technologies and evaluation for Central and Eastern European languages” in conjunction with International Conference RANPL’2009. Borovec, Bulgaria, 17 September 2009. 1-8. ISBN 978-954-452-008-3; 1. Dimitrov, L., Koseska, V., Garabík, R., Erjavec, T., Iomdin, L., Shyrovkov, V. MONDILEX Towards the Research in Slavic Lexicography. In: International Journal Cognitive Studies/Études Cognitives. Vol. 10, SOW, Warsaw, 2010. 147-162. ISSN 2080-7147; Dimitrova, L., Koseska–Toszeva, V., Roszko, D., Roszko, R. Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus. In: International Journal Cognitive Studies/Études Cognitives. Vol. 10, SOW, Warsaw, 2010. 217-240. ISSN 2080-7147; L.,Dimitrova, V. Koseska, R. Garabík, T. Erjavec, L. Iomdin, V. Shyrovkov. Conceptual Scheme for a Research Infrastructure Supporting Digital Resources in Slavic Lexicography. Sofia, Demetra Publ. House, 2010. ISBN 978-954-8986-33-5)

The bilingual texts juxtaposed in the Corpus have been specially selected and are oriented towards contemporary problems of semantics and comparative linguistics. The first Polish-Bulgarian parallel electronic collection of experimental comparative research (i.e. the Bulgarian-Polish Comparative Corpus by L. Dimitrova and V. Koseska) contributed to the creation of experimental Bulgarian-Polish electronic dictionary (L. Dimitrova, V. Koseska, J. Satoła-Staškowiak).

⁷Institute of Slavic Studies, Polish Academy of Sciences.

⁸Institute of Mathematics and Informatics, Bulgarian Academy of Sciences.

We already know that the precondition of the selection of texts will be, first of all, finding the largest possible number of texts in Bulgarian on account of a small number of their translations into Polish and definitely larger into Russian. The largest quantity of texts existing in Polish or Bulgarian can be found in the third language — Russian.

Bulgarian, Russian and Polish differ typologically — Bulgarian is analytic whereas Polish and Russian are synthetic languages. Our parallel corpus should have compatible annotation which at the same time takes into account the features that distinguish the selected languages.

The texts in our corpus will be juxtaposed on the level of a sentence. Naturally, every text entered into the corpus will have its equivalent in the two remaining languages. The language of the original text is not specifically defined here and will rather depend on the availability of the material and as a result the direction of translation will undergo changes.

2. The test version of the corpus (the current state of development)

Work on the project began in autumn 2010. Within a few weeks we managed to start the first version of Corpus. Here, we introduce the current state of works and we will explain certain issues relating to Corpus' development.

Corpus already runs online as a test version, but is not yet publicly accessible. Corpus' software is written in the PHP language, with the use of the MySQL database managing system. PHP, as well as MySQL are free environments, which can be downloaded by anyone and used in their own projects. Corpus can be run either on Apache server or other servers supporting PHP and MySQL technologies.

From user side Corpus has been tested in various browsers, such as Internet Explorer 6 for Windows, Opera for Windows, Opera Mobile (Symbian) and Firefox for Windows and Ubuntu Linux. With regard to Windows, Corpus works at its optimum when using the Firefox browser. At the moment, Corpus consists of two parts: the first of these is prose and the second includes works of poetry. Technically, both sub-corpus look alike, however, the texts in the prose sub-corpus are aligned at the sentence level, and in the poetry sub-corpus at the verse level. Let's focus first on the prose corpus and after that we will say a few words about the poetry corpus.

The first version of the prose sub-corpus is not very sizeable and is very much intended as a test version. It contains the first few chapters of Bulgakov's *Master and Margarita*, a hundred pages of Zeromski's *Ashes (Popioły)* as well as books translated by Violetta Koseska-Toszewa from Bulgarian into Polish.

The Corpus interface, as seen by users, appears as shown in Fig. 1 below.

Figure 1 The Corpus Interface

While searching for language elements in Corpus, it is essential to correctly set up the language in which a given expression is searched. Searching is possible in all three languages of Corpus: Bulgarian, Russian and Polish.

Statystyka:
Liczba wyników: 12
Liczba zdań, w których je znaleziono: 12

Русский	Polski	Български
Высокий тенор Берлиоза разносился в пустынной аллее, и по мере того, как Михаил Александрович забирался в дебри, в которые может забираться, не рискуя свернуть себе шею, лишь очень образованный человек. - поэт узнавал все больше и больше интересного и полезного и про египетского Осириса, благостного бога и сына Неба и Земли, и про финикийского бога Фаммуза, и про Мардука, и даже про менее известного грозного бога Виллилуци, которого весьма почитали некогда ацтеки в Мексике.	Wysoki tenor Berlioz rozlegał się w pustej alei i im dalej Michał Aleksandrowicz zapuszczał się w gąszcz, w który bez ryzyka skręcenia karku zapuścić się może tylko człowiek niezmiernie wykształcony - tym więcej ciekawych i pożytecznych wiadomości zdobywał poeta. Dowiedział się na przykład o łaskawym bogu egipskim Osyrysie, synu Nieba i Ziemi, o sumeryjskim bogu Tammuzie, o Marduku i nawet o mniej znanym groźnym bogu Hutlilopochin, którego niedawno wiele powołał meksykańscy Aztekowie.	Високият тенор на Берлиоз кънтеше в пустата алея и колкото повече навлизаше Михаил Александрович в тези дебри, в които без риск да си строши главата можеше да навлезе само много образован човек, толкова повече интересни и полезни неща научаваше поетът и за благия египетски бог Осирис, син на Небето и Земята, и за финикийския бог Тамуз, и за Мардук, и дори за по-неизвестния страшен бог Уиллилуци, дълбоко почитан някога от ацтеките в Мексико.
Вы полностью повторили мысль беспокойного старика Иммануила по этому поводу.	Pan dokładnie powtórzył pogląd nieokiełzanego staruszka Immanuela w tej materii.	Вие повторихте изцяло мисълта на неспокойния старец Имануил по този въпрос.
Удивление выразилось на лице секретаря, сорбирующего над низким столом и записывающего показания.	Zdziwienie odmalowało się na twarzy sekretarza, który pochylony nad niskim stołem spisywał zeznania.	Учудване се изписа на лицето на секретаря, който записваше показанията, приведен над ниската масичка.
Круто, исподобя Пилат буравил глазами арестанта, и в эти глаза уже не было мутн, в них появились всем знакомые искры.	Płot surowo przyglądał się aresztowanemu spość łba, w jego oczach już było już mgiełki, pojawiły się w nich dobrze wszystkim znane iskry.	Остро, изпод вежди, Пилат пронизваше с поглед арестанта и в очите му вече нямаше мътилка, в тях се появили познатите на всички искри.
- Не знаешь ли ты таких, - продолжал Пилат, не сводя глаз с арестанта, - некого Димоса, другого - Гестаса и третьего - Вар-равана?	- Czy znany ci jest - ciągnął Płot nie spuszczaając wzroku z więźnia - niejaki Dimos albo Gestas, albo może Bar Rawan?	- А да познаваш — продължи Пилат, без да сваля очи от арестанта, — някой си Димас, и друг един, Гестас, и трети — Вар-Раван?
Тот вздрогнул, обернулся, но успокоил себя мыслью, что его имя и отчество известны профессору также из книги-небудь.	Redaktor drgnął, odwrócił głowę, ale sam siebie uspokoił, że	Оня трепна, обърна се, но се успокои с мисълта, че и

Figure 2 Query results

Figure 3 Using the ‘translit’ option

Zapytanie: ros. *говори*

Statystyka:
Liczba wyników: 66
Liczba zdań, w których je znaleziono: 58

Русский	Polski	Bulgarski
- Нет ни одной восточной религии, - говорила Берлиоз, - в которой, как правило непорочная дева не произвела бы на свет бога	- Nie ma takiej wschodniej religii - mówił Berioz - w której dziewczina nie przrodziła boga	— Няма нито една източна религия — говореше Берлиоз, — в която някоя непорочна дева да не е родила бог
- Ты, Иван, - говорила Берлиоз, - очень хорошо и сатирически изобразил, например, рождение Иисуса, сына божия, но соль-то в том, что еще до Иисуса родился еще ряд смылов божиих, как, скажем, фригийский Атис, коротко же говоря, ни один из них не родился и никого не было, в том числе и Иисуса, и необходимо, чтобы ты, вместо рождения и, скажем, прихода волхвов, описал нелепые слухи об этом рождении...	- A ty - mówił Berioz do poety - bardzo dobrze i odpowiednio satyrycznie pokazałeś, na przykład, narodziny Jezusa, syna bożego, ale dowieść polega na tym, że jeszcze przed Jezusem narodziło się całe mnóstwo synów bożych, jak powiedzmy fenicki Adonis, frygijski Atis, perski Mitra. A tymczasem, krótko mówiąc, żaden z nich się w ogóle nie narodził, żaden z nich nie istniał, nie istniał także i Jezus. Musisz koniecznie zamiast narodzin Jezusa czy też, powiedzmy, hołdu trzech królów, opisać nonsensowne wieści rozpowszechniane o tym hołdzie	— Ти, Иване — говореше Берлиоз, — много дубаво и сатирично си описал например раждането на Исус, сина божии, но работата е там, че още преди Исус са били родени редица синове божии, като, да речем, фригийският Атис, какъто казвам, нито един от тях никога не се е раждал, нито е съществувал, изключително Исус, и ти вместо раждането му или, да речем, поклонението на вълхвите, трябва да опишеш нелепите слухове за това раждане
- Извините меня, пожалуйста, - заговорил подошедший с иностранным акцентом, но не коверкая слов, - что я, не будучи знаком, позволяю себе... но предмет вашей ученой беседы настолько интересен, что...	- Proszę mi wybaczyć - zaczął nieznanymy. Mówił z cudzoziemskim akcentem, ale słów nie kaleczył - że nie będąc znajomym panów ośmielam się... ale przedmiot naukowej dyskusji panów jest tak interesujący, iż...	— Моля да ме извините — заговори приближилният с чуждестранен акцент, но без да изопачава думите, — че без да се познаваме, си позволявам... но темата на вашния научен разговор е толкова интересна, че...
- Нет, вы не ослышались, - учтиво ответил Берлиоз, - именно это я и говорила	- Tak jest, nie przesłyszał się pan - grzecznie odpowiedział Berioz - To właśnie powiedziałem	— Не, не грешите — учтиво отвърна Берлиоз, — тъкмо това казах
И недаром Шидлер говорила , что кантовские рассуждения по этому вопросу могут удовлетворить только рабов, а Штраус просто смеялся над этим	Tę nie na darmo Schiller powiada, że rozważania Kanta na ten temat mogą zadowolnić tylko ludzi o duszach niewolników, a Strauss je po prostu śmieje się	И не случайно Шилер казва, че Кантовите разсъждения по този въпрос могат да задоволят само роби, а Штраус просто се посмее от това разсъждение

Figure 4 The result of the search for “govoril” (with ‘translit’ option enabled) in Russian texts

In the current version of Corpus there are 4 modes of searching.

Firstly, one can search for any sequence of letters, from one letter up to a sentence-like sequence. For instance, searching in Polish texts for the sequence “*znan*” outputs all the words containing this sequence of letters (as well as the sentences containing these words): *znanym*, *nieokiełznanego*, *znane*, *znany*, *nieznanym*.

As we said before, it is possible to search in all three languages. For users’ convenience, the option ‘Translit’ has been added. Thanks to this option, the users can type Cyrillic words by using the Latin alphabet and keyboard layout (see Fig. 3, 4).

In the prose sub-corpus, as we said before, the texts are aligned at the level of the sentence. In case when a sentence in one language is represented by two or more sentences in another language, these two or more sentences are still treated as equivalent of that sentence. This can be illustrated by the following examples:

Русский	Polski	Български
<i>Иван ажнул, глянул вдаль и увидел ненавистного неизвестного. Тот был уже у выхода в Патриарший переулок, и притом не один.</i>	<i>Iwan jęknął, spojrzął w dal i przy samym wyjściu zobaczył zniepokojonego nieznanego, który już nie był sam.</i>	<i>Иван ажна, погледна в далечината и видя омразния непознат. Она беше вече на изхода откъм улица Патриаршая и при това не беше сам.</i>
<i>В руке Иван Николаевич нес зажженную венчальную свечу. Правая щека Ивана Николаевича была свежее изодрана.</i>	<i>Iwan trzymał w dłoni zapaloną świeczkę, a na jego prawym policzku widniało świeże zadrapanie.</i>	<i>Иван Николаевич носеше запалена венчална свещ. Дясната буза на Иван Николаевич беше прясно разкървавена.</i>

In both examples, one Polish composed sentence will be translated into two sentences in Russian or Bulgarian.

Compare also the following example (a query: Bulgarian „глава”):

Русский	Polski	Български
<i>В этой голове гудел тяжелый колокол, между глазами яблоками и закрытыми веками проплывали коричневые пятна с огненно-зеленым ободком, и в довершение всего тошнило, причем казалось, что тошнота эта связана со звуками какого-то назойливого патефона.</i>	<i>W głowie tej kołysał się olbrzymi dzwon, pod powiekami przepływały brązowe plamy z ognistozieloną obwódką. Na domiar wszystkiego Stiope mdliło, przy czym wydawało mu się, że mdłości te wiążą się w jakiś sposób z natrętnymi dźwiękami patefonu.</i>	<i>В тази глава бучеше тежка камбана, между очните му ябълки и затворените клепачи плуваха кафяви петна с огненозелени краища, а на всичкото отгоре му се повръщаше и му се струваше, че това има връзка със звуците на някакъв натрапчив грамофон.</i>

There is only one Russian or Bulgarian composite sentence joined by an appropriate conjunction (Russian *u*, Bulgarian *a*), while in Polish two sentences are used (without any conjunction).

In the case that a given sentence does not have an equivalent in one or both languages of the corpus, this sentence is removed. This seems justified by the fact that the parallel texts in Corpus are a set of corresponding units, and if a certain sentence is not translated (or is omitted) it means that the equivalent in another language cannot be found. However, this is not often the case.

At this point it is worth mentioning that there are many programs used for sentence alignment (such as Hunalign, Bitext2tmx, etc.). This software is able to align texts in two languages (the source language and the translation), however, not in three languages. As far as we know there is no any free software that would allow for the alignment of three languages, therefore we used a special alignment procedure. Bulgarian and then Russian texts were aligned to Polish by the use of free software called TextAlign (see: Textalign). In addition, we were using a PHP program, created by us, which allows to merge aligned pairs of texts (Bulgarian-Polish, Russian-Polish), make further alignment (Bulgarian-Polish-Polish), and then add the result to Corpus.

It is also possible to search in Corpus for whole words or for sequences of words. The word is understood here, simply, as a sequence of letters between spaces, between a space and a punctuation mark and between a punctuation mark and a space. An example of a search in the mode ‘whole word’ (a query: Polish “*pan*”) is presented below (Fig. 5):

Русский	Polski	Български
- Если я не ослышался, вы изволили говорить, что Иисуса не было на свете? - спросил иностранец, обращая к Берлиозу свой левый зеленый глаз.	- Jeśli dobrze usłyszałem, to był pan łaskaw stwierdzić, że Jezus w ogóle nie istniał? - zapytał kierując na Berliozą swoje lewe zielone oko.	— Ако не греша, вие благоволихте да кажете, че Иисус изобщо не е съществувал на този свят, нали? — попита чужденецът, обръщайки към Берлиоз зеленото си ляво око.
- Нет, вы не ослышались, - учтиво ответил Берлиоз, - именно это я и говорил.	- Tak jest, nie przesłyszał się pan - grzecznie odpowiedział Berlioz - To właśnie powiedziałem.	— Не, не грешите — учтиво отвърна Берлиоз, — тъкмо това казах.
- А вы соглашались с вашим собеседником? - осведомился неизвестный, повернувшись вправо к Бездомному.	- A pan zgodził się z kolegą? - zainteresował się nieznanym i odwrócił się w prawo, do Bezdomnego.	— А вие бихте съгласни с вашия събеседник? — осведоми се непознатият и се обърна надясно към Бездомния.
Тут иностранец отложил такую штуку, встал и пожал изумленному редактору руку, произнес при этом слова - Позвольте вас поблагодарить от всей души!	Wtedy cudzoziemiec wykonał taki numer: wstał, uściłinal zdumionemu redaktorowi dłoń; powiedział, co następuje - Niech pan pozwoli, że mu z całego serca podziękuję.	Товага чужденецът направи следния номер: стана, стисна ръка на изумения редактор и каза. — Позвовете, да ви благодаря от все сърце!
- За что это вы его благодарите? - заморгав, осведомился Бездомный.	- Za co mu pan dziękuję? - mrugając osłupia zapytał Bezdomny.	— Но за какво му благодарите? — запремита Бездомния.
Ведь согласитесь, что в области разума никакого доказательства существования бога быть не может.	Przyzna pan chyba, że w kategoriach rozumu nie można przeprowadzić żadnego dowodu na istnienie Boga.	Та съгласете се, че в областта на разума не може да има никакво доказателство за съществуването на бога.
Вы полностью повторили мысль беспокойного старика Иммануила по этому поводу.	Pan dokładnie powtórzył pogląd niepokojącego staruszka Immanuela w tej materii.	Вие повторихте изцяло мисълта на неспокойния старец Иммануил по този въпрос.
Ведь говорил я ему тогда за завтраком «Вы, профессор, ваши слова, что-то несладкое придумали! Оно, может, и умно, но только непонятно. Над вами потешаться будут».	Przecież mówiłem mu wtedy, przy śniadaniu: «Jak pan tam, profesorze, sobie chce, ale wymyślił pan coś, co się kupić nie trzyma. Może to i mądre, ale zbyt skomplikowane. Wymyślęj panasza».	Нами му казах тогава на закуска. „Ще прочавате, професоре, но сте измислили нещо много неспокойно. То може и да е умно, но е страшно объркано. Ще ви се смеят.“

Figure 5 A search in the mode ‘a whole word’

The same request, but specified as a ‘sequence of letters’, gives different results (see Fig. 6).

Русский	Polski	Български
Второй - плечистый, рыжеватый, выхристый молодой человек в замоченной на затылок клетчатой кепке - был в ковбойке, жевал белых броек и в черных тапочках.	Drug, rudawy, barczysty, kudyślaty młody człowiek w zwinętej na cieniu krzasiastej cyklistówce i krzasiastej koszuli, szedł w wymiętych białych spodniach i czarnych płóciennych pau toch.	Вторият, широкоплещест младеж с буйна червеникава коса и жилият на тила кариран каскет, носеше спортна риза, смачкан бѣл панталон и черни платненки обувки.
Тут ужас до того овладел Берлиозом, что он закрыл глаза.	Berkoza opa owało takie przerażenie, że aż zamknął oczy.	Такавъ ужас обзе Берлиоз, че той затвори очи.
Он был в дорогом сером костюме, в заграничных, в цвет костюма, туфлях.	Miał na sobie drogi szary garnitur i dobre pod kolor pa nowe.	Носеше сѣл сив костюм и чуждестранни обувки сѣцѣя цвѣт.
Тут Бездомный сделал попытку прекратить замучившую его шпату, задержав дыхание, отчего икнул мучительнее и громче, и в этот же момент Берлиоз прервал свою речь, потому что иностранец вдруг поднялся и направился к писателю.	W tym momencie Bездомny próbował opa ować, czkając go ciekawą i wstrzymał oddech, na skutek czego sznął jeszcze boleśniej i głuszej i w tej chwili Berkoza przerwał swój wykład, ponieważ cudzoziemiec wstał nagle i zbliżył się do pisarza.	В този миг Бездомни задрѣжа дѣх, да спре хълцането, което ужасно го мъчеше, но само изхълца още по — мъчително и по-силно и в сѣция момент Берлиоз мѣкна, защото чужденецът изведнѣж стана и се запѣти към писателите.
- Извините меня, пожалуйста, - заговорил подошедший с иностранным акцентом, но не коверкал слов, - что я, не будучи знаком, позволю себе... но предмет вашей ученой беседы настолько интересен, что...	- Proszę mi wybaczyć - zaczął nieznamy. Mówił z cudzoziemskim akcentem, ale słów nie kalceżył - że nie będąc znanym pa nowi tłumaczł się... ale przedmiot naukowej dyskusji pa now jest tak interesujący, iż...	- Моля да ме извините — заговори приближилит се с чуждестранен акцент, но без да изопачава думите, — че без да се познаваме, си позволювам..., но темата на вашия научен разговор е толкова интересна, че...
- Разрешите мне присесть? - вежливо попросил иностранец, и приятеля как-то несколько радевинувшись, иностранец ловко усеялся между ними и тотчас вступил в разговор.	- Pa nowie pozwól, że się przysiądam? - uprzejmie zaprosił cudzoziemiec i przykładał jakoby mimo woli się rozsiadł, cudzoziemiec ловко усеял się pomiędzy nich i natychmiast włączył się do rozmowy.	- Ще ми позволите ли да поседна? — учтиво попита чужденецът и приятелите нѣкак неволно се дръпнаха един от друг, а чужденецът първо се настани между тях и веднага се намеси в разговора.

Figure 6 Searching for a sequence of letters.

It is also possible to search for all the grammatical forms of a given lexeme (only for Russian or Polish). For example, requesting “*ty*” in this mode one can search for Polish word *ty* as well as other grammatical forms (cases) of it, such as *tobą*, *tobie* etc.; a query ‘*iść*’ will return *idzie*, *szedł*, *szli*, *szła* etc. Examples (Fig. 7–9):

Русский	Polski	Български
Прямо, и выйдем куда надо.	- Тędy, proszę. Prosto, a traficie, gdzie należy.	— Оттук, ако обичате!
Пропустив мимо себя все три вагона, кот вскопши на заднюю дугу последнего, лапой вцепился в какую-то кишу, вышло из стени, и утяти, сжоним, таким образом, тривеники.	Przeczekal, aż minę go wszystkie trzy wagony, po czym wykoszył na tylny zderzak ostatniego, łapą objął sterzącą nad zderzakiem gumową rurę i pojechał. Zaoszczędzwszy w ten sposób dziesięć kopiejek.	Той изчака да отминат и трите вагона и скопи на задния бѣфер на последния, вкопчи се в ѣвжмавъ маркуч, който се подвѣшаваше от трамвая, и офейка, спестив по този начин десет копейки.
А ресторант заки своей обычной ночной жизнью и жил бы ею до заката, то есть до четырех часов утра, если бы не произошло нечто, уже совершенно из ряду выходящее и поражающее ресторанных гостей гораздо больше, чем известие о гибели Берлиоза.	Restauracja zaś zaczęła swoje zwykłe nocne życie i kontynuowałaby je aż do zamknięcia, to znaczy do czwartej rano, gdyby nie zaszło coś zupełnie już niezwykłego, coś, co na gościach restauracji zrobiło znacznie większe wrażenie niż wieść o śmierci Berkoza.	А ресторантът отново заживя своя нормален нощен живот и щеше да си го живее до затварянето, тоест до четири часа сутринта, ако не беше се случило нещо вече съвсем необичайно, което изуми посетителите на ресторанта много повече от вѣстта за гибелта на Берлиоз.
Швейцар, вышел в этот момент из дверей ресторанный вешники во двор, чтобы покурить, затопил папиросу и двинулся было к привидению с явной целью претрадить ему доступ в ресторан, но почему-то не сделал этого и остался, глуповато улыбаясь.	Portier, który właśnie wyszedł na papierosa przed drzwi restauracyjnej szatni, zapałal niedopałek i ruszył w kierunku zjawy, najwyraźniej chcąc jej zagrozić drogę do restauracji, ale z jakichś nieznamych powodów odstąpił od tego zamiaru i stanął uśmiechając się głupkowato.	Портиерът, който тъкмо беше изшъл от гардероба на ресторанта в градината да пуши, смачка с крак цигарата и поведе да приближи привидението с явно намерение да му претради пътя към ресторанта, но кой знае защо не го стори и се спря с глупава усмивка.
Через четверть часа чрезвычайно пораженная публика не только в ресторане, но и на самом бульваре и в окнах домов, вышло в сад ресторана, видела, как из ворот Грибодова Пантелей, швейцар, милиционер, официант и поэт Рюкин выносили спеленатого, как куклу, молодого человека, который, заливаясь слезами, плакался, норовя поспать именно в Рюкина, давился слезами и кричал: - Сволочи!	W kwadrans potem niezmiernie porażona publiczność nie tylko w restauracji, ale także na bulwarze i w oknach sąsiadujących z ogrodem restauracji domów, patrzyła, jak Panielej, portier, milicjant, kelner oraz poeta Ruchin wywiozą z bramy Gribojedowa człowieka, który jak niemowlę młodego człowieka, który zalewał się łzami, płł, szczególnie starając się trafić w Ruchina, i krzyczał na cały bulwar: - Drazgi! ... Drazgi! ...	След четвърт час потресените присъстващи не само от ресторанта, но и на булеварда, и по прозорците на къщите, които гледаха към градината на ресторанта, видѣха как през вратата на Грибодов Пантелей, портиерът, един милиционер, един келнер и поетът Рюкин изнасят повит като кукла младеж, облян в слъзи, който плюсе, мъчейки се да улечи именно Рюкин, и крещи: — Мръсници!
Слово «что» он произнес дискантом, «а» - басом, а «утождо» у него совсем не вышло .	Słowo «czego» zostało wypowiedziane falsetem, «a» - basem, a «uczucie» - w ogóle nie wydostało się na świat bazy.	Думата «кой» произнесе с дискант, «сте» — басово, а от «вие» изобщо нищо не се получи.

Figure 7 The query Rus. “*vyjti*” (“*выхити*”)

Zapytanie: pol. *pracować*

Statystyka:
Liczba wyników: 4
Liczba zdań, w których je znaleziono: 4

Русский	Polski	Български
Мне всегда как-то лучше работается за городом, в особенности весной.	Zawsze jakoś lepiej mi się <i>pracuje</i> za miastem, zwłaszcza na wiosnę.	Не знам защо, но винаги по ми върви работата извън града, особено пролетно време.
Вот только дама, которую Степа хотел поцеловать, осталась неразъясенной... черт ее знает, кто она... кажется, в радио служит, а может быть, и нет.	Tylko dama, którą Siopa chciał pocałować, pozostawała w dalszym ciągu niewyjaśniona... diabli wiedzą, co za jedna... zdaje się, że <i>pracuje</i> w radio, a zresztą może i nie...	Само дамата, която Степа се беше опитал да целуне, оставаше неизяснена... дявол я знае коя е... май работи в радиото, пък може и да не работи там.
- И ничего на делаешь?	- Zatem nie <i>pracujesz</i> ?	- И нищо ли не работиш?
- Придворные мыши никогда не работают.	- Dworzanie nigdy nie <i>pracują</i> .	- Придворните мишки никога не работят.

Zakończono

Figure 8 The query Pol. “*pracować*”

Polski	Русский	Български
nie dobiegając do turникiet, tak i ostanął na niej.	не добежав до турникета, так и остался на ней.	да дотича до турникета и както се строполи на една пейка, тъй си и остана.
Zawienela padająca so stołowi posuda, zakrzykali żonkami.	Rozległ się brzęk spadających ze stołów naczyń i krzyki	Завънтяха падащи от масите съдове, жените закрещяха.
Opłochem się szwajcar usłyszał z werandy uszanie, bój posudy i żeński kłótni.	Wtedy na wprost oszalały portier uderzył jakowejś jęki dobiegające z werandy, szczęk tłuczonych naczyń, krzyki	Обезумелият портier чу откъм верандата глъч на сбиване, шум от строшени съдове и женски писъци.
On uszedł powrócić na bok, beśnym dźwiękiem w tenże mgnienie podniósł nogi i, powierając się, rozglądał się na niego z nieudrżymą siłą, zupełnie białe od ukaślenia żonkami-wagowojazów i se alupowku.	Michał Aleksandrowicz zdążył przekroczyć się na bok, wielkim ruchem, w teje sekunde podciągnął nogi pod brzuch, odwrócił się i zobaczył niepowstrzymanie pędzącą na niego zbieżną ze zgrozą twarz prowadzącej tramwaj <i>kobiety</i> i jej czerwona opaskę.	Успя да се извърти на хълбок и в същия миг с бясно движение да свие крака към корема, но когато се обърна, съзря понеслото се към него с неударжима сила съвсем бляло от укаś лице на жена ватманка и алената ѝ забрадка.
Однако нежданно воле него столкнулись две женщины, одна из них, остроносая и протолопая, закричала над самым ухом поэта другой женщине так: - Аннушка, наша Аннушка!	Nagle niespodziewanie tuż przed nim wpały na siebie dwie <i>kobiety</i> , jedna z nich ostronośa i rozczochrana, tuż nad uchem poety krzyknęła do drugiej, co następuje: - Annuska, to nasza Annuska!	Неочаквано пред него се сблъскаха две жени и едната, остроноса и чорлава, закрещя на другата над самото ухо на поета: — Аннушка, нашата Аннушка!
Потеря одного из преследуемых, Иван сосредоточил свое внимание на коте и видел, как этот странный кот подошел к подножке моторного вагона «А», стоящего на остановке, нагло отсадил ввинчившую женщину, уцепился за поручень и даже сделал попытку всунуть кондукторше тринижен через открытое по случаю дулуты окно.	Iwan, zgubiwszy jednego ze ściganych, całą uwagę skoncentrował na kocurze i zobaczył, że dzwiny ów kot podzielił do drzwi wagonu motorowego lin «A», który stał na przystanku, beczelnie odepchnął wrzeszczącą <i>kobieta</i> , chwycił za poręcz i nawet wykonał próbę wtęczenia konduktorkę dziesiątką przez otwarte z powodu upału okno.	След като загуби единия от преследваните, Иван съсредоточи вниманието си върху котаракъа и видя как този страшен котарак се приближи към стъпалото на първия вагон на трамвай „А“, който беше на спирката, как изтика нагло една жена, която извинча, как се хвана за дражката и дори се опита да пробута на кондукторката десет копейки през отворения поради жегата прозорец.
Какие же меры мы приняли, чтобы поймать этого убийцу? - тут врач повернулся и бросил взгляд женщине в белом халате, сидящей за столом в стороне.	A co pan zrobił, aby schwycić tego mordercę? - w tym momencie lekarz odwrócił się i spojrzął na siedzącą przy stoliku pod ścianą <i>kobietę</i> w białym fartuchu.	И какви мерки взехте, за да заловите този убиец? — при тези думи лекарят се обърна и хвърли поглед към жената с бела престилка, седнала в страни до една маса.
Пьянствуют, вступают в связи с женщинами, используя свое положение, ни черта не делают, да и делать ничего не могут, потому что ничего не смыслят в том, что им	Pija, wykorzystując swoje stanowisko spią z <i>kobietami</i> , ni cholery nie robią, zresztą nawet nie mogą nic robić, bo nie	Пиянстват, използват положението си за връзки с жени, нищо не вършат, пък и за нищо не ги бива, защото бълк не разбират от

Zakończono

Figure 9 The query Pol. “*kobieta*”

The following programs were used in order to enable searching for all available forms of a lexeme:

- TaKIPi (tagger for Polish language by IPI PAN, see Piasecki 2007);
- Mystem (tagger for Russian language by Yandex, see Mystem).

So far, these programs were of our interest only as lemmatisers, i.e. tools, which specified the basic lexical form of words in the analyzed texts. Before these texts were included in Corpus they were processed by the above programs. Subsequently, the files with the results were processed by specific PHP scripts, which produced a form of morphological dictionaries, containing pairs ‘the form as appearing in the text — basic lexical form of this word’. Thanks to this procedure, it was possible to work out a preliminary version of an algorithm for searching for all the forms of a given lexeme.

It is important to emphasize that, in its present shape, Corpus is not annotated morphosyntactically, which is why the problem of the homonymic forms of different lexemes, which are pronounced equally, has not been resolved. For example, the user searches for the forms of the Russian lexeme *веко* (‘eyelid’), but the result will show lots of words, which are the forms of the Russian noun *век* (‘100 years’). These different expressions have convergent forms in the oblique cases (for example, the instrumental plural: *веками*).

Русский	Polski	Български
- Да, по этой пригласили. - подтвердил профессор и пояснил: - Тут в государственной библиотеке обнаружены подлинные рукописи чернокнижника Герберта Аврелиаского, десятого века , так вот требуется, чтобы я их разобрал.	- Tak, jako specjalistę - potwierdził profesor i wyjaśnił: - W naszej bibliotece narodowej znaleziono oryginalne rękopisy Herberta z Aulais, z dziesiątego wieku. Poproszono mnie, żebym je odcyfrował.	— Да, във връзка с нея съм поканен — потвърди професорът и обясни: — Във висшата държавна библиотека са открити оригинални ръкописи на чернокнижника Херберт Аврелиаски от десети век. Та трябва да ги разчета.
Вспушшее веко приподняло, подернутый дымкой страдания глаз уставился на арестованного.	Uniosła się opuchnięta powieka, zasłute mgiełką cierpienia oko wpatrzyło się w aresztowanego.	Подпухналият клепач се поддигна, забуленото в пелената на страданиято око се втрени в арестувания.
Лишь только группа, выйдя из сада на площадь, поднялась на обширный царский над площадью каменный помост, Пилат, оглядываясь сквозь прищуренные веки , разоблался в обстановке.	Skoro tylko cała grupa, wyszedłszy z ogrodów na plac, weszła na rozległy, górujący nad placem kamienny pomost, Płat zmrużył oczy i rozglądał się badając sytuację.	Щом групата излезе от градината на площада и се изкачи на широкия, извисяващ се над площада каменен подиум, Пилат, който се оглеждаше през полуотворени клепачи, веднага се ориентира в обстановката.
Он постараясь выкинуть котыбы последние и для этого разделил савишися веки левого глаза.	Postanowił wykinąć przynajmniej ten ostatni problem i w tym celu rozdzielił depione powieki lewego oka.	Помъчи се да си изкачи поне последното и с тази цел разпели слепените клепачи на лявото си око.
Степа разделил слепенные веки и увидел, что отражается в трюме в виде человека с торчащими в разные стороны волосами, с опущен, покрытою черной щетиною физиономией, с заплаканными глазами, в грязной сорочке с воротником и галстуком, в кальсонах и в носках.	Rozdzielił depione powieki i ujrzał w lustrze człowieka ze sterzącymi na wszystkie strony włosami, z opuchniętą, pokrytą szarą szczecinią fiziozomią, z zapłakniętym osupłym. Czdowiek ów miał na sobie brudną koszulę, krawat, kalesony i skarpetki.	Степа разпели слепените си клепачи и видя, че се отразява в огледалото като човек с торчащи на всички посоки коса, с подпухнала, обрасла с черна четина физиономия, плунали в глътна очи, с мръсна риза и вратовързка, по наполеони и по чорапи.
Под веками у него вспыхнул зеленый огонь, от него загорелся мот, и над толпою полетели зримые арамейские слова - Четверо преступников, арестованных в Ершалаиме за убийства, подстрекательства к мятежу и оскорбление законов и веры, приговорены к позорной казни - повешению на кресте.	Zapłonął pod jego powiekami zielony ogień, mógł zajął się od tego płomienia i poszybowały ponad tłumem ochryple aramejskie słowa - Czterech przestępców zatrzymanych w Jerozolimie za morderstwa, podburzanie do buntu, obrażenie praw i religii skazanych zostało na hańbiącą śmierć przez powieszenie na krzyżu.	Под клепачите му думна зелен огън, от който моят му пламна и над тълпата полетяха зримки арамейски думи: — Четиримата престъпници, арестувани в Йерусалим за убийства, подстрекателства към бунт и оскорбление на законите и вярата, са осъдени на позорна смърт — окачване на крест.

Figure 10 Some gramatical forms of the Russian lexeme *век* among the results of the search for the Russian lexeme *веко*.

Certainly it is not possible to avoid such homonymity unless there is morpho-syntactic annotation added.

The fourth Corpus searching mode is the searching with the use of regular expressions. This mode gives users many possibilities. For example, they can browse for expressions (words) that share the same root, but have different prefixes (after listing the prefixes). A request “(za/przy/wy/pod/od)szedł” will result in, for example: *przyszędł, wyszedł, podszędł, odszedł* (see Fig. 11).

Figure 11 The use of the fourth Corpus searching mode (regular expressions).

Figure 12 An example of a search in the poetry subcorpus.

It is necessary to mention the poetry sub-corpus as well. Currently it contains the poem of Nikolay Zydarov *Мунка малката маймунка*, translated into Polish by Koseska-Toszewa. Technically, the poetry sub-corpus works the same way as the prose sub-corpus, so we won't repeat the rules of operating which are stated

above. However, sub-corpus differ in their content. Poetry translations constitute special — both theoretical and practical — problem. In translations of poetry works a slight change of meaning may be necessary for the rhyme and rhythm's sake. Therefore the translation and the original in case of a poetical work may differ more significantly than is the case with prose work.

In Bulgarian original sentence (see Fig. 12 above) — *Па захвърли с всичка сила орех и към крокодила. На жирафа каза с яд: — Ех че смешно дълговат!* — the nuts are thrown at the crocodile, in Polish translation it is coco nuts that are thrown at the crocodile, while in Russian the nuts are thrown at an antelope.

Conclusions

A semantics team of the Institute of Slavic Studies of the Polish Academy of Sciences directed by Violetta Koseska began works on creation of the Bulgarian-Polish-Russian Parallel Corpus. It would be the first publicly available parallel corpus of these specific languages. The planned corpus will be based on the material from one period (the 20th century) and will have a synchronous character.

An introductory version of the corpus was already developed and is presented in this paper. Naturally, we did not touch upon many of the problems or issues related with the corpus and its creation. The primary and most complicated task, for the future, would be an executing the morphological annotation for all three languages. Various technical improvements, such as dividing results into pages and sorting options, will be introduced gradually.

References

- Dimitrova, Koseska 2009:** Ludmila Dimitrova, Violetta Koseska-Toszewa, *Bulgarian-Polish Corpus*, Cognitive Studies — Etudes Cognitives 9, Slawistyczny Ośrodek Wydawniczy = SOW Publishing House, Warsaw 2009, 133–141.
- Dimitrova, Koseska, Roszko 2009a:** Ludmila Dimitrova, Violetta Koseska, Danuta Roszko, Roman Roszko, *Bulgarian-Polish-Lithuanian Corpus — Current Development*, [In:] Cristina Vertan, Stelios Piperidis, Elena Paskaleva, Milena Slavcheva (Eds.), *International Workshop. Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages held in conjunction with The International conference RANLP — 2009, Proceedings*, Borovets 2009: 1–8.
- Dimitrova, Koseska, Roszko 2009b:** Ludmila Dimitrova, Violetta Koseska, Danuta Roszko, Roman Roszko, *Bulgarian-Polish-Lithuanian Corpus — Problems of Development and Annotation*, [in:] Tomaž Erjavec (Ed.), *Research Infrastructure for Digital Lexicography, Information Society 2009, MONDILEX Fifth Open Workshop, Ljubljana, Slovenia, October 14–15, 2009, Proceedings*, Ljubljana, Jožef Stefan Institute, 2009, p. 72–86.
- Dimitrova, Koseska, Roszko 2010:** Ludmila Dimitrova, Violetta Koseska, Danuta Roszko, Roman Roszko, *Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus)*, Cognitive Studies - Etudes Cognitives 10, Slawistyczny Ośrodek Wydawniczy = SOW Publishing House, Warsaw 2010, p. 217–240.
- Mystem:** <http://company.yandex.ru/technology/mystem/>

- Piasecki 2007:** Piasecki M., *Polish Tagger TaKIPI: Rule Based Construction and Optimisation*. Task Quarterly 2007, 11, p. 151–167.
- TextAlign:** <http://mt2007-cat.ru/>
- Woliński, 2003:** Woliński M., *System znaczników morfosyntaktycznych w korpusie IPI PAN*. Polonica XII, 2003, p. 39–55.