MAKSIM DUŠKIN¹, JOANNA SATOŁA-STAŚKOWIAK²

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 it.

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 Mondilex, 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 PAN, NKJP, Bul-Pol Korpus). The authors of

---

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

3 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 and 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. Satola-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.

---

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-corporuses 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 Bulgarian-Polish-Russian Parallel Corpus

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

![Corpus Interface](image1)

**Figure 1** The Corpus Interface

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

![Query results](image2)

**Figure 2** Query results
Figure 3 Using the ‘translit’ option

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): \textit{znanym}, \textit{nieokiełznanego}, \textit{znanego}, \textit{znany}, \textit{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:

<table>
<thead>
<tr>
<th>Русский</th>
<th>Polski</th>
<th>Бългаски</th>
</tr>
</thead>
<tbody>
<tr>
<td>Иван ахнул, глянул вдаль и увидел ненавистного неизвестного. Тот был уже у выхода в Патриарший переулок, и при том не один.</td>
<td>Iwan jęknął, spojrzał w dal i przy samym wyjściu zobaczył znienawidzonego nieznajomego, który już nie był sam.</td>
<td>Иван ахна, погледна в далечината и видя омрачен непознат. Оня беше вече на изхода от улица Патриаршва и при това не беше сам.</td>
</tr>
</tbody>
</table>

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“):

<table>
<thead>
<tr>
<th>Русский</th>
<th>Polski</th>
<th>Бългаски</th>
</tr>
</thead>
<tbody>
<tr>
<td>В этой голове гудел тяжелый колокол, между глазными 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.</td>
<td>W głowie tej kołysał się olbrzymi dzwon, pod powiekami przepywali brązowe plamy z ognistozieloną obwódką. Na domiar wszystkiego Stiopę mdliło, przy czym wydawało mu się, że mdłości te wiązać się w jaki sposób z natrętnymi dźwiękami patefonu.</td>
<td>В тази глава бучеше тежка камбана, между очни те му ябълки и затворени клепачи плуваха кафяви петна с огнено-зелени краища, а на всичкото отгоре му се повръщаше и му се струваше, че това има връзка със звуци те на някакъв натрапчив грамофон.</td>
</tr>
</tbody>
</table>
There is only one Russian or Bulgarian composite sentence joined by an appropriate conjunction (Russian и, Bulgarian а), 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, Bitex2tmx, 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):

![Figure 5 A search in the mode ‘a whole word’](image-url)
The same request, but specified as a 'sequence of letters', gives different results (see Fig. 6).

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 toby as well as other grammatical forms (cases) of it, such as toby, tobie etc.; a query 'iść' will return idzie, szedł, szli, szła etc. Examples (Fig. 7–9):
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).

Certainly it is not possible to avoid such homonymity unless there is morphosyntactic 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)szedl” will result in, for example: przyszedł, wyszedł, podszedł, 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
The Bulgarian-Polish-Russian Parallel Corpus above. However, sub-corporuses 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 смисълъ дълговрат! — 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


Mystem: http://company.yandex.ru/technology/mystem/
