COGNITIVE STUDIES | ÉTUDES COGNITIVES, 15: 57–65
Warsaw 2015

DOI: 10.11649/cs.2015.005

JOANNA SATOŁA-STAŚKOWIAK

Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland
joannasat@op.pl

ON THE BENEFITS OF FOREIGN LANGUAGE LEARNING BASED ON PARALLEL LANGUAGE CORPUS

Abstract

A recently observed strong interest in language corpora, which can be defined as a collection of texts in an electronic format, as well as my work within the European Project Clarin on 'The Parallel Polish-Bulgarian-Russian Corpus' became the reason for writing the text concerning the use of the parallel language corpus for learning a foreign language. The article discusses the benefits resulting from the use of such a corpus in learning a foreign language, describes selected corpus language tools supporting the learning process as well as indicates some threats arising from the wrong use of the corpus.

Keywords: language corpora; foreign language learning; translation; concordance tools; formation of electronic resources; The Parallel Polish-Bulgarian-Russian Corpus

1. Language corpora, because of the dynamic development of corpus linguistics and computer technologies, are extremely important research material, used among others in applied linguistics (cf. Hunston, 2002) and lexicography (cf. Ooi, 1998). In my view, language corpora should play a bigger role in developing language skills than to date. In recent years many researchers (Maia, 2003, p. 52), academic teachers and not a very numerous group of foreign language teachers, who attended conferences dedicated to the issue, have taken to these tools.

The incorporation of corpus methodology and introduction of corpus tools to foreign languages learning should become a fact. The creative search for words, the possibility of contrasting units, constructions and utterances in language A with language B and others would probably lead to many individual discoveries and give answers to many questions which are essential at a given stage of learning. The questions which so far almost only foreign language teachers have been able to answer.

In learning a foreign language both corpora of one language as well as parallel corpora can be useful. Annotated and unedited (electronic) text collections are important since both types of material provide linguistic data worth looking at.

The benefits of teaching Polish, Bulgarian or Russian languages on the basis of parallel corpus are merely signalled in the article. 'The Parallel Polish-Bulgarian-Russian Corpus',[1] to which the author of the article with the team of corpus linguistics and semantics had contributed, became a fundamental source and, among other lexicographic aids, helped in writing the first volume of 'The Contemporary Bulgarian-Polish Dictionary' (cf. Satoła-Staśkowiak & Koseska-Toszewa, 2014). It is also the basis of a trilingual Russian-Bulgarian-Polish dictionary which is being compiled (Kisiel, Koseska, Sosnowski).

1.1. 'The Parallel Polish-Bulgarian-Russian Corpus' contains text collections exceeding 6 million forms (common value for three languages). The collected materials are made up of pieces of fiction, technical instructions, legal and other texts in Polish, Bulgarian and Russian. The texts were obtained in three different ways: 1. Texts based on free access, 2. Texts with the obtained author's licence, 3. Copyright-exempt texts.

1.2. The Corpus has been designed from the theoretical-methodological side in such a way so as to maintain the proportions of the above mentioned kinds of texts. The possibilities of its use are diverse. It can be helpful, among other things, in writing dictionaries, grammar textbooks, scientific articles, specialist thematic sub-corpora and in teaching the Polish, Bulgarian or Russian language.

2. In learning a foreign language any language corpus whose resources contain a representative number of examples together with their language context can be helpful. It is good if it is an annotated collection which provides for the possibility of asking detailed questions and enables the user to find information which is of interest to them. From the corpus user's point of view both morphosyntactic and stylistic markers, arising from metadata placed with every text (containing information on the author, translator, the time of edition and the place of editing the work) can be essential. The trilingual corpus discussed here in its first version contains solely semantic annotation which concerns only some examples chosen from the corpus (about 1/10 examples from the whole corpus) as well as metadata selected according to the pattern discussed above.

A lack of annotation of lexical units in the corpus is not an obstacle to using its resources. Such a corpus can be a practical tool supporting learning a foreign language. Especially as, like in the case of 'The parallel Polish-Bulgarian-Russian Corpus, on the level of a sentence three languages are collated in it simultaneously and the user of the parallel corpus should be very familiar with at least one of them.

[1]Parallel Polish-Bulgarian-Russian Corpus is being compiled by the Team of Computational Linguistics and Semantics of the Institute of Slavonic Studies, Polish Academy of Sciences (PAS) (V. Koseska-Toszewa, A. Kisiel, J. Satoła-Staśkowiak, W. Sosnowski,) taking part in a European project Clarin (Common Language Resources and Technology Infrastructure).The founders of Clarin ERIC are Austria, Bulgaria, the Czech Republic, Denmark, Estonia, Germany, Holland and Poland. The main aim of the project is to combine resources and language tools for the European languages into one common uniformed network which is to become an important tool of work for academics from broadly understood humanistic branches of science.

The benefits resulting from using material which is not annotated were described by M. Wilkinson (n.d.) (the article was published on the Internet without the date of publication) listing, among other things: a possibility of confirming oneself in intuitive decisions, affirming or changing decisions (on the choice of unit and construction) based on other sources such as e.g. dictionaries, obtaining information about possible collocations, broadening knowledge on the subject of patterns in the target language, learning how to use new expressions.[2]

Another advantage of learning a language on the basis of language corpora is the fact that thanks to them lexical units are presented in a broader context, enabling advanced learners to make their own basic linguistic analysis.

2.1. In the future intelligent concordance programs supporting advanced analysis of texts included in 'The Parallel Polish-Bulgarian-Russian Corpus' will be used for following specific units and constructions. At present only ordinary filters facilitate searching for a definite word or derivational or morphological element in the text.

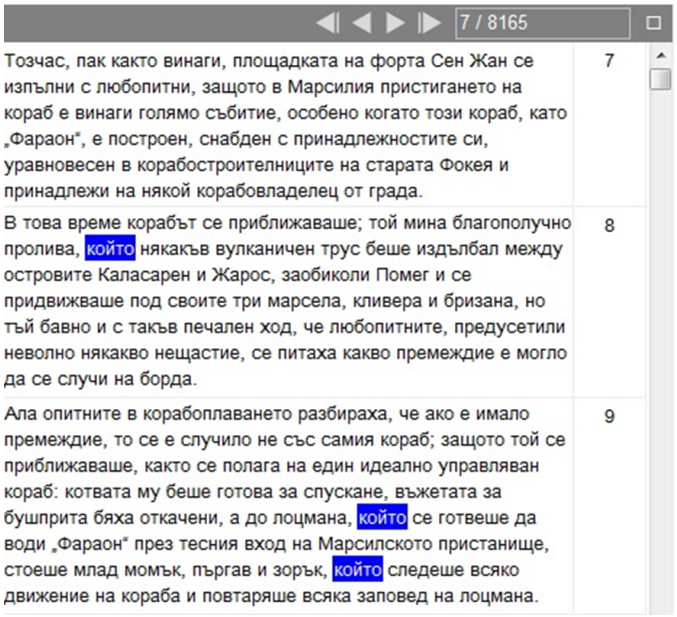

Figure 1. presents the operations of a filter indicating a key word in context in 'The Parallel Polish-Bulgarian-Russian Corpus'

The electronic tools used in the corpus show the frequency of occurrence of specific units or constructions thus aiding learners (which is especially important for

[2]Such learning has particular importance when, as in the case of the Polish and Bulgarian language there is only one complete Bulgarian-Polish and Polish-Bulgarian dictionary by F. Sławski and S. Radewa, which was published in Poland the last time in 1987 and 1988 and which does not contain contemporary (mentally embedded in the 21st century) language material and the corpus of both languages is a current lexicographical source.

them) in creating their own lists, e.g. put in alphabetical order, forming synonymous or antonymous groups, groups of neologisms or archaisms as well as other units important to the individual user. The filters also support, in accordance with the adopted methodology, the segmentation of language material (in the case of the language corpus discussed here) into sentences.

Table 1:

Polski	**Български**	**Русский**
Zważywszy, że wszelkie działania oraz inicjatywy dotyczące działalności przeciwko handlowi ludźmi muszą być nie dyskryminujące i brać pod uwagę równość płci, jak również podejście uwzględniające prawa dziecka;	Като имат предвид, че всички действия или инициативи, насочени срещу трафика на хора, не трябва да бъдат дискриминационни, а да отчитат равенството между половете, както и подхода за закрила на правата на децата;	Считая, что любая деятельность или инициатива в области борьбы с торговлей людьми должны осуществляться без какой-либо дискриминации и учитывать равенство между женщинами и мужчинами, а также подход, в основе которого лежат права ребенка;
24 lutego roku 1815 strażnik morski z Notre Dame de la Garde zasygnalizował przybycie trójmasztowca „Faraon", powracającego ze Smyrny przez Triest i Neapol.	На 24 февруари 1815 година дежурният наблюдател на Нотър Дам дьо ла Гард възвести пристигането на тримачтовия кораб „Фараон", който идваше от Смирна, Триест и Неапол.	Двадцать седьмого февраля 1815 года дозорный НотрДам де-ла-Гард дал знать о приближении трехмачтового корабля "Фараон", идущего из Смирны, Триеста и Неаполя.

3. The parallel Polish-Bulgarian-Russian corpus (further: PPBRC) shows the user collocations of definite units and the number of positions they determine in three different languages, thus marking the dissimilarity of the collated systems, or, conversely, they confirm a common way of explication of chosen constructions.

Thanks to following examples in three languages together with their contexts PPBRC can recognise the semantic value of the analysed units. The user has access to knowledge concerning each of the three languages, which is not in any way restricted on account of difficulty or easiness of understanding the collated material. It makes the corpus material different from educational material (e.g. from textbooks for learning a foreign language) of individual languages, intended for a student with elementary, intermediate or advanced knowledge of a specific language. The third language collated in the corpus (it can be both Russian or Bulgarian as well as Polish) constitutes a kind of additional linguistic background.

PPBRC provides the user with information on word order, exchangeability of individual lexical elements and changes of meaning that some exchanges in lexical

constructions carry with them (Bogusławski, 1976, 1994).

(1) Pol. Wiele czasu spędziłem z dorosłymi.
Bulg. Живял съм много при възрастните.
Rus. Я долго жил среди взрослых.

(2) Pol. Ludzie zajmują na Ziemi bardzo mało miejsca.
Bulg. Хората заемат съвсем малко място на Земята.
Rus. Люди занимают на Земле не так уж много места.

PPBRC allows looking at the equivalents of such lexical units as: Polish *na domiar* (on top of all that) , *na skutek* (as a result), *na zawsze* (for ever), *chodzi o* (the point is), *na dodatek* (in addition), *do diabła* (to hell):

(3) Pol. Proszę pana — mówił dalej Edmund — wiem, że nie może mnie pan zwolnić z więzienia *na skutek własnej decyzji.*
Bulg. Знам, господине — продължи Дантес, — че не можете да ме извадите оттук *по собствено решение.*
Rus. Я знаю, — продолжал Дантес, — я знаю, что вы не можете освободить меня *своей властью.*

(4) Pol. — *Do diabła!* — wykrzyknął Albert.
Bulg. — *Дявол да го вземе!* — каза Албер.
Rus. — *Черт возьми!* — сказал Альбер.

PPBRC aids the analysis of forms whose formal equivalents exist in language A but do not exist or are becoming extinct in langugae B or C, cf. e.g. the perfective participle in the Polish and Bulgarian language (Satoła-Staśkowiak, 2009).

(5) Pol. — Proszę! Oto badacz! — wykrzyknął, *ujrzawszy* Małego Księcia.
Bulg. Я, гледай! Един изследовател! — извика той, когато *забеляза* малкия принц.

(7) Pol. Villefort, *wróciwszy*, zamknął za sobą drzwi, ale zachwiał się — on teraz — kiedy wszedł do salonu.
Bulg. Вилфор *влезе*, затвори вратата, но щом стигна в салона, краката му се подкосиха.

PPBRC imparts information on inflected, semantic and even pragmatic qualities. It illustrates the frequency of using anglosemantisms or internationalisms in each of the collated languages (Satoła-Staśkowiak, 2014).

(8) Pol. Zdając sobie sprawę, że skuteczna walka z *cyberprzestępczością* wymaga zwiększonej, szybkiej i dobrze funkcjonującej współpracy międzynarodowej w sprawach karnych;
Bulg. Като преценяват, че ефективната борба срещу престъпността в *кибернетичното пространство* изисква мащабно, бързо и ефикасно международно сътрудничество в наказателно правната област;

Rus. Полагая, что для эффективной борьбы против *компьютерных преступлений* требуется более широкое, оперативное и хорошо отлаженное международное сотрудничество в области уголовного права;

(9) Pol. Świadome głębokich zmian dokonanych na skutek *digitalizacji, konwergencji* i trwającej *globalizacji sieci informatycznych*;

Bulg. Като осъзнават дълбоките изменения, предизвикани от *въвеждането на информационните технологии, от конвергенцията* и от постоянната *глобализация на компютърните мрежи.*

Rus. Сознавая глубокие перемены, вызванные внедрением *цифровых технологий, объединением* и продолжающейся *глобализацией компьютерных сетей*;

PPBRC enables the user to follow equivalents of a given unit in a specific kind of texts, e.g. in legal texts. Definite kinds of texts introduce restrictions concerning the use of a given unit (despite the fact that there is a whole collection of synonymous units). Cf. (examples come from 5 legal texts in three langauges collated in the corpus i.e. about 70 000 words) For Polish *jednak* (however) (29) — Bulgarian *все пак* (4), *независимо* (6), *освен* (7), *въпреки това* (9), lack of translation (3) — Russian *однако* (15), *только* (4), lack of translation (10).

Polish *jednakże* (after all) (15) — Bulgarian *независимо* (6), *независимо от* (4), *освен ако* (1), lack of translation (4) — Russian *однако* (13), lack of translation (2).

And others, coming from legal texts, most often used equivalents of units of the kind: Polish *na podstawie* (based on) — Bulgarian *на основата на* — Russian *на том основании*; Polish *na mocy* (on the strength of sth) — Bulgarian *по силата на* — Russian *согласно*; Polish na piśmie (in writing) — Bulgarian *в писемна форма* — Russian *в письменном виде*; Polish *na dowód* (as proof) — Bulgarian *в увериениа* — Russian *в удостверение*; Polish *w szczególności* (in particular) — Bulgarian *в частност* — Russian *в частности*; Polish *w przypadku / w przypadkach* (in the case of / in cases of) — Bulgarian *в случай на / в случаите* — Russian *в случае / в случаях*; Polish *w odniesieniu do* (with respect to) — Bulgarian *по отношение* — Russian *в отношении*.

(10) Pol. *W odniesieniu do* państwa, które podpisało protokół, a następnie go ratyfikuje, przyjmie lub zatwierdzi, wchodzi on w życie po upływie 90 dni od daty złożenia przez nie dokumentu ratyfikacyjnego, przyjęcia lub zatwierdzenia.

Bulg. *По отношение на* подписала държава, която ратифицира, приеме или утвърди протокола впоследствие, той влиза в сила 90 дни след датата на депозиране на документа й за ратификация, приемане или утвърждаване.

Rus. *В отношении* подписавшего государства, которое ратифицирует, примет или одобрит Протокол после этого, он вступает в силу через 90 дней после сдачи на хранение его ратификационной грамоты или документа о принятии или одобрении.

The observed frequency of the use of definite units in a specific kind of texts in PPBRC can also be confronted (just to be sure) with other corpora the user is familiar with, e.g., for Polish, with the National Corpus of Polish Language.

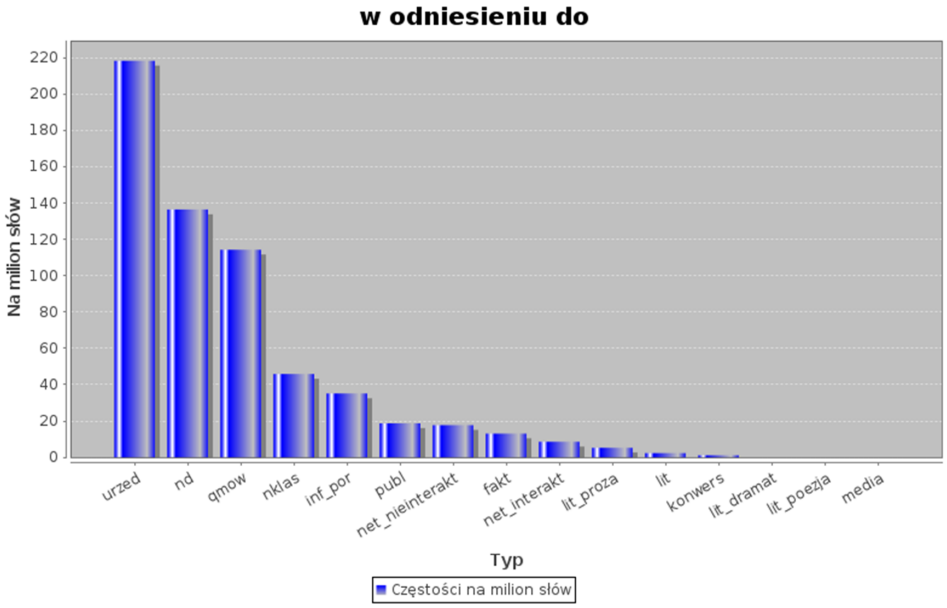

Figure 2. `http://www.nkjp.uni.lodz.pl/index_adv.jsp`

3.1. The corpus allows following many linguistic phenomena in quite a short time. Perhaps it will become a fairly serious source of information at least half as important as classic electronic or printed dictionary. It has great importance in the case of learning a Bulgarian language since the only 'complete' Polish-Bulgarian and Bulgarian-Polish dictionaries by S. Radewa and F. Sławski were written 26 and 27 years ago and the lack of newer lexicographical titles describing comprehensively general contemporary Bulgarian language is still noticeable. An added advantage is the possibility of becoming acquainted with translation techniques (cf. Satoła-Staśkowiak, 2014).

4. It seems that the possibilities that the corpora give together with the programs supporting them — looking for key words, collocations or suitability of expressions or constructions far outweigh potential threats that some researchers indicate in the literature on the subject (cf. Ball, 1997; Stewart, 2000). After all, these threats are described mainly in connection with the translator's work and the possibilities that the corpus translation memory brings. The translator, backed by advanced tools, can treat the solutions suggested by translation memory as authority, forgetting about their own creative input into the translated work. (Satoła-Staśkowiak, 2014). However, this is not a rule.

Instead, a multitude of translation solutions observed in the corpus is the best incentive to learn a language and understand subtle semantic differences between examples. It is important that no corpus constitutes the only source of information on a language. The corpus described here has to be treated as an aid in the process of education, which in conjunction with other existing sources (e.g. dictionaries, grammar textbooks) will ensure deeper and more reliable knowledge.

In the near future (at the end of the year 2015) everyone interested in Polish, Bulgarian or Russian language will be able to use PPBRC. At this time it will be available on the Internet. Unlimited access to the corpus and the fact that the digitalized and parallelized text resources will be consistently expanded will allow verifying knowledge about its actual usefulness in learning the three Slavonic languages discussed in the article.

References

Ball, C. (1997). *Concordances and corpora*. Georgetown University. Retrieved October 1, 2014, from `http://www.georgetown.edu/faculty/ballc/corpora/tutorial1.html`

Benis, M. (1999). *Translation memory from O to R*. Retrieved 1 October 2015, from `http://utkl.ff.cuni.cz/~rosen/VYUKA/MT/tm-review01.htm`

Bogusławski, A. (1976). O zasadach rejestracji jednostek języka. *Poradnik Językowy*, (8), 356–364.

Bogusławski, A. (1994). Obiekty leksykograficzne a jednostki języka. In A. Bogusławski, *Sprawy słowa* [Word matters] (pp. 115–124). Warszawa: Veda.

Friedbichler, I. & Friedbichler, M. (1997). *The potential of domain-specific target-language corpora for the translator's workbench*. Paper presented at the first international conference on Corpus Use and Learning to Translate, Bertinoro, 14–15 November 1997.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Maia, B. (2003). Some languages are more equal than others: Training translators in terminology and information retrieval using comparable and parallel corpora. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 43–53). Manchester: St Jerome.

Satoła-Staśkowiak, J. (2009). Translating into something that does not exist... — literary ways of translating Polish sentences with uninflected perfect participles into the Bulgarian language. *Cognitive Studies | Études cognitives*, *9*, 211–221.

Satoła-Staśkowiak, J. (2013a). Contemporary contrastive studies of Polish, Bulgarian and Russian neologisms versus language corpora. *Cognitive Studies | Études cognitives*, *13*, 143–160. `http://doi.org/10.11649/cs.2013.009`

Satoła-Staśkowiak, J. (2013b). Neologizmy bułgarskie, polskie i rosyjskie w ujęciu konfrontatywnym. In D. Blagoeva, S. Kolkovska, & M. Lishkova (Eds.), *Problemi na neologiiata v slavianskite ezitsi* (pp. 21–30). Sofiia: BAN.

Satoła-Staśkowiak, J. (2013c). Polskie i bułgarskie neologizmy znaczeniowe. In D. Blagoeva, S. Kolkovska, & M. Lishkova (Eds.), *Problemi na neologiiata v slavianskite ezitsi* (pp. 218–230). Sofiia: BAN.

Satoła-Staśkowiak, J. (2014a). Different aspects of neosemantization on the example of the Polish and Bulgarian language. *Cognitive Studies | Études cognitives*, *14*, 183–191. `http://dx.doi.org/10.11649/cs.2014.015`

Satoła-Staśkowiak, J. (2014b). Edukacja przyszłych tłumaczy w oparciu o korpusy językowe. In *Prikladna linhvistyka ta linhvistychni tekhnolohiï, MegaLing-2013* (pp. 211–223). Kyïv: NAN Ukraïny, Ukr. movno-inform. fond.

Satoła-Staśkowiak, J. & Koseska-Toszewa, V. (2014): *Współczesny słownik bułgarsko-polski*. (A. Kisiel, Ed.). Warszawa: Slawistyczny Ośrodek Wydawniczy.

Stewart, D. (2000). *Supplying native speaker intuitions or normalising translation? Translating into English as a foreign language with the British National Corpus*. Manchester: Research Models in Translation Studies.

Varantola, K. (2003). Translators and disposable corpora. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 55–70). Manchester: St Jerome.

Wilkinson, M. (n.d.). Using a specialized corpus to improve translation quality. *Translation Journal*, *9*(3). Retrieved 10 October 2015, from `http://translationjournal.net/journal/33corpus.htm`

Acknowledgment

This work was supported by a grant from CLARIN (Common Language Resources and Technology Infrastructure).

The author declares that she has no competing interests.

Publisher: Institute of Slavic Studies, PAS, University of Silesia & The Slavic Foundation