THE PARALLEL POLISH-BULGARIAN-RUSSIAN CORPUS: PROBLEMS AND SOLUTIONS

Wojciech Sosnowski
The Institute of Slavic Studies of
the Polish Academy of Sciences

Abstract

The parallel Polish-Bulgarian-Russian corpus we are currently developing as part of CLARIN-PL framework will become an essential tool for translators producing both traditional and digital translations. The electronic tools developed within the project facilitate fast search for and retrieval of multilingual equivalents of lexemes, phrases and sentences. Selected sentences and texts have been semantically annotated for the quantification of nomen, time and aspect. Our definition of equivalent stems from the contemporary contrastive linguistics theory. The guiding principle in the construction of the corpus was to proceed from meaning to form; the principle was first introduced in Koseska-Toszewa (2006).

During our work on the Polish-Bulgarian-Russian corpus, we have come across a number of issues, which we regard as characteristic of multilingual corpora: (1) the selection and procurement of texts, (2) the development of computer tools used for the construction of the corpus, (3) multilingual equivalence, and (4) semantic annotation.

Multilingual corpora have proved to be exceptionally helpful in language teaching, traditional and digital lexicography, as well as traditional and digital translations. The usefulness of multilingual corpora in each of these areas will be demonstrated through example corpus queries.

1. INTRODUCTION

The parallel corpora we are currently developing as part of CLARIN-PL framework (a Polish-Bulgarian-Russian corpus and a Polish-Lithuanian corpus) will become essential tools for translators producing both traditional and digital translations. In linguistics, parallel corpora will enable provide large amounts of data for the study of language and its evolution. Parallel corpora are also useful in language teaching, sociology, cultural studies as well as other fields related to linguistics and information technology. In the 2000s, many countries developed their national corpora, e.g. Poland (the so-called "one-million" National Corpus of Polish), Bulgaria (Bulgarian National Corpus, http://search.dcl.bas.bg/) and Russia (Russian National Corpus, http://www.ruscorpora.ru). Although the above corpora have proved to be valuable tools for linguists studying these languages in isolation, they were of little use to scholars working in contrastive linguistics, lexicography, translation studies and language teaching.

2. PARALLEL CORPORA IN CLARIN-PL

The Department of Corpus Linguistics and Semantics of the Polish Academy of Sciences has been developing a parallel Polish-Bulgarian-Russian corpus, which is to be incorporated into the CLARIN framework[106]. Our corpus will become the first multilingual corpus of Slavonic languages. It has been our priority to develop a multilingual corpus, because monolingual and even bilingual corpora are inadequate tools for comparative linguistics.

The European Union aims to make its ubiquitous digital market truly multilingual. The ubiquitous digital market strategy must address all issues that relate to multilingualism in order to ensure that EU offers equal opportunities for speakers of each of EU's official languages. Nevertheless, the language barrier still remains the main barrier to a truly integrated European economy and society. In order to overcome this barrier, we have been working on a number of corpora as part of CLARIN-PL: a Polish-Bulgarian-Russian corpus and a Polish-Lithuanian corpus. These corpora will bridge the gap in Slavonic and Balto-Slavonic digital linguistic resources and will help provide accurate translations of digital and conventional texts.

As soon as we began our work on the parallel corpora, a number of problems emerged that were specific to multilingual corpora. The remainder of this section will give an overview of the issues we encountered and the solutions that we chose to address them.

2.1. Selecting the languages

We have chosen Polish, Bulgarian and Russian because they are representative of the West, South and East Slavonic group respectively. The languages exhibit different structures: synthetic (Polish and Russian) and analytic (Bulgarian). They also employ different writing systems: the Latin script (Polish) and the Cyrillic script (Russian and Bulgarian).

2.2. Selecting the texts

The first version of the corpus will contain 6 million words — 2 million words for each language. We plan to add another 2 million in the second stage of the project. The aim of the planned expansion (the second stage) is to make our corpus a medium-sized corpus, which will enable researchers to conduct novel types of studies with the use of the corpus. The selection of texts for the corpus was based on the following criteria:

1) Texts must come from different styles, genres and registers (general language, languages for special purposes)

2) Texts must come from different sources (the original text in one of the corpus's languages or a translation from a different language into all three languages of the corpus)

3) Texts must come from different historical periods

4) Every text must exhibit a high standard of language correctness (critically acclaimed translations, canonical literary texts)

---

106 Common Language Resources and Technology Infrastructure is a project granted the status of ERIC (European Research Infrastructure Consortium) by the European Commission in February, 2012. CLARIN was founded by eight countries: Austria, Bulgaria, the Czech Republic, Denmark, Estonia, Germany, the Netherlands and Poland. CLARIN is part of the ESFRI (European Roadmap for Research Infrastructures, European Strategy Forum on Research Infrastructures). The project's primary aim is to combine language tools and resources for multiple European languages into one unified network, which will become an important research tool for scholars in arts, humanities and social sciences.

Eventually, we have included multiple text genres in the corpus: literary texts from the 19th, 20th and 21st century, instruction manuals and technical documentation, legal texts, as well as other types of documents. The table below presents some example texts included in the corpus:

Name	Word count
Additional Protocol to the European Convention on Mutual Assistance in Criminal Matters	3371
Antoine de Saint-Exupery, *The Little Prince*	35228
European Convention on Transfrontier Television	14621
Amendments to the European Convention on Transfrontier Television	14328
European Convention the Archaelogical Heritage	6843
Convention on the Protection of Children against Sexual Exploitation and Sexual Abuse	21749
Convention on the Recognition of Qualifications concerning Higher Education in the European Region	14937
Council of Europe Convention on Action against Trafficking in Human Beings	21563
Statute of the Council of Europe	8330
Paulo Coelho, *Eleven Minutes*	18946
Statute of the Council of Europe	34613
European Convention on Recognition and Enforcement of Decisions concerning Custody of Children and on Restoration of Custody of Children	9690
Universal Declaration of Human Rights	4442
European Cultural Convention	2859
Council of Europe Convention on preventing and combating violence against women and domestic violence	29668
European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment	6481
European Convention on Mutual Assistance in Criminal Matters	8884
Additional Protocol to the European Convention on Extradition	2964
Joseph Conrad, *Lord Jim*	92307
Stefan Żeromski, *Ashes*	68132
Angel Wagenstein, *Far From Toledo*	184 421
Kyoto Protocol	21 257
Paulo Coelho, *The Alchemist*	17 636
Alexandre Dumas, *The Count of Monte Cristo*	417 620
Paulo Coelho, *The Witch of Portobello*	20 096

As we can see, beside literary texts the corpus also incorporates a large number of documents produced by international institutions, e.g. Council of Europe treaties and official EU documents.

2.3. Obtaining texts

The texts in the corpus come from three sources: (1) open source publications; (2) copyrighted documents for which we have obtained licenses[107] and (3) public domain texts (i.e. texts whose intellectual property rights have expired or have been forfeited). The search engine in the final version of the corpus will only display as much text as it is allowed by the right to quote. Every text will be annotated with metadata, which will also be displayed by the search engine. Yet another problem that we encountered while working on the corpus was that some texts have not yet been converted to an electronic format and therefore we had to digitise them manually. So as to obtain the most accurate version possible, after every phase of digitising a text was proofread and edited.

2.4. Developing the corpus with computer tools

The first step in developing the corpus was to choose a computer application that would enable us to align three languages in parallel. When we began the work on our corpus, it became clear that there was no application that would allow us to split large texts in three different languages in parallel. Eventually, we decided to use NOVA Text Aligner. NOVA Text Aligner is a tool designed to make manual text alignment as easy and simple as possible. There are automated paragraph/sentence alignment tools but there is one thing that they all have in common – they are not 100% accurate (and they can not be due to the nature of the task they are supposed to do). So this means that in the end you'll have to go trough the whole text yourself and check it and correct it (http://www.supernova-soft.com/wpsite/products/text-aligner/). First, we would align Polish and Bulgarian texts and afterwards we would supplement them with the third language. While aligning the texts, we found that the sentence-level equivalence was very difficult to achieve.

3. MULTILINGUAL EQUIVALENTS IN CONTRASTIVE STUDIES

The definition of equivalence that we follow in our research derives from the contemporary semantic theory and contrastive studies of natural languages developed in the multi-volume *Gramatyka konfrontatywna bułgarsko-polska* [further referred to as: GKBP] (Koseska-Toszewa and Gargov, 1990; Koseska-Toszewa, 2006; Koseska-Toszewa, Korytkowska and Roszko, 2007). GKBP is the first contrastive grammar in the world that makes use of an intermediate semantic interlanguage. Using a semantic interlanguage to compare multiple languages provides an innovative solution for contrastive studies and diverges from traditional principles of applied contrastive studies. Traditionally, the comparison between two (or more) languages relied heavily on the primary language of description. In consequence, it was always incomplete and could also be misleading, if not grossly inaccurate.

In theoretical contrastive studies, the analysis of language data proceeds from meaning to form. This stands in contrast to traditional contrastive grammars, which tend to depart from a form in one language and then proceed to a form in another language. The above procedure – outlined in GKBP – enabled us to treat the data from every language as equal.

[107] The development of a model licence agreement took approximately one year.

Equivalence or the lack of equivalence is a widely debated phenomenon in linguistics:

> Equivalence (or lack thereof) is a marginal phenomenon, if comparative studies take under consideration only one language. The notion of equivalence, on the other hand, plays a crucial role in contrastive lexicology. Accordingly, the notion of equivalence in lexicology concentrates on the language system, therefore it is relatively vague. On the basis of designation lies a polysemic understanding of the linguistic sign. Consequently, an element of the lexicon can have several values, i.e. meanings. When comparing an element from the language A with another element in the language B, generally the denotative relationship is the basis for such a comparison. Thus, there is an equivalence, which is usually called semantic equivalence with the provisos that, firstly, the number of sememes in language A is the same as in language B (and thus they have the same value), and, secondly, their denotation (paired sememes) is the same. (Jaskot, in press).

The table below presents a selection of equivalent sentences that we have encountered:

Agent okrętowy nie potrzebuje zdawać żadnych egzaminów, ale musi posiadać zdolność abstrakcyjnego myślenia i umieć wykazać je w praktyce	На морския кларк не са нужни никакви изпити, но той трябва да се отличава с абстрактно умение и да го проявява на практика.	Морскому клерку не нужно сдавать никаких экзаменов, но предполагается, что он должен отличаться сноровкой и проявляет ее на практике.

Joseph Conrad, *Lord Jim*

Tu postanowił spędzić noc. Wprowadził swoje owce przez rozpadającą się bramę i zagrodził wejście deskami tak, by w nocy nie mogły się wymknąć.	Реши да пренощува тук. Вкара овцете през разнебитената порта и я залости с няколко дъски така, че да не могат да избягат.	Он решил заночевать там, загнал через обветшавшую дверь своих овец и обломками досок закрыл выход, чтобы стадо не выбралось наружу.

Paulo Coelho, *The Alchemist*

Szum ów dobiegał z dołu, wznosił się ku stropom, ku twarzy prokuratora. A za plecami Piłata, za skrzydłem pałacu grały larum trąbki, słychać było stamtąd ociężały chrzęst setek nóg i pobrzękiwanie żelastwa. Procurator zrozumiał, że to już wymarsz piechoty rzymskiej, która spełniając jego rozkaz, udaje się na straszną dla buntowników i zbójców przedśmiertelną defiladę.	Този шум се надигаше отдолу към нозете и лицето на прокуратора, а зад гърба му, там, отвъд крилата на двореца, се чуваха тревожните сигнали на тръбите, тежкото скърцане на стотици стъпки, звън на желязо. Прокураторът разбра, че римската пехота вече тръгва, съгласно неговата заповед, за страшния за бунтарите и разбойниците предсмъртен парад.	Этот шум поднимался снизу к ногам и в лицо прокуратору. А за спиной у него, там, за крыльями дворца, слышались тревожные трубные сигналы, тяжкий хруст сотен ног, железное бряцание, – тут прокуратор понял, что римская пехота уже выходит, согласно его приказу, стремясь на страшный для бунтовщиков и разбойников предсмертный парад.

Mikhail Bulgakov, *The Master and Margarita*

4. SEMANTIC ANNOTATION

We are currently working on the semantic annotation of 2000 sentences in the parallel Polish-Bulgarian-Russian corpus and the Polish-Lithuanian corpus. The preliminary annotation (i.e. the 2000 sentences we are working on at the moment) must be performed manually. Once it is completed, it will serve as a basis of an automatic tagger. Koseska-Toszewa & Roszko (2015) developed an innovative semantic annotation scheme, which can be applied to entire sentences in multilingual parallel dictionaries. Instead of choosing a number of separate sentences and annotating them, we will annotate longer fragments of texts. The semantic annotation scheme outlined in Koseska-Toszewa & Roszko (2015) will

make contrastive studies of natural languages easier and, in consequence, facilitate more efficient manual and automatic translations. Below, I will present some examples of the semantic annotation scheme at work.

5. POSSIBLE IMPLEMENTATIONS OF THE CORPUS

Our corpus constitutes a comprehensive resource for scholars developing on multi-lingual dictionaries or conducting studies comparing multiple languages (e.g. Bulgarian-Polish and Russian-Polish contrastive grammar, Polish-Lithuanian contrastive grammar, Balto-Slavonic contrastive studies. The target group of our corpora are linguistics; stylisticians; translators (e.g. to investigate translation strategies employed in the works available in the corpus); scholars (e.g. for studies on terminologies and lexical equivalence); students of literary studies (for comparative research), cultural studies (to study the forms culturemes take in different languages), sociology (the texts we included in our corpus are a reflection of the social processes that took place in their respective periods), political studies, history, intercultural communication or anthropology.

- Searching for equivalents necessary for synchronic contrastive studies.

- Developing translation memories (TMs) based on contemporary lexis; these translation memories can be later incorporated into translators' own translation memories. Translation memories should be developed in a widely recognised format (e.g. TMX), which can be imported into the most popular CAT application suites. TMs significantly reduce the amount of time and labour translators and teachers have to spend on their tasks. More importantly, they also enable automated database search, which ensures high stylistic and terminological coherence of texts produced by translators and teachers.

- Quantitative studies (frequencies of word types and tokens as well as syntactic structures and contexts they appear in).

- Data necessary for the construction of grammatical models of languages.

- Research on intercomprehension[108]: it provides data for the construction of exercises that aim at the activation of the passive knowledge of cognate languages. In the area of Slavonic languages, exercises of this type are quite an innovation; they are of paramount importance, especially taking into consideration the fact that Slavonic languages form a significant part of the linguistic landscape in the EU and, what is more, they are quite closely related to each other.

- Investigating translation strategies: comparing the lexical and grammatical constructions in different languages used for the expression of similar semantic content; studying how different languages convey phraseological units, culturemes and non-equivalent lexis; stylistic and terminological coherence, etc.

- Teaching of first and second languages.

- Studies of text-level equivalence of culturemes.

- Quotation search.

- And many more.

[108] Cf. The European Intercomprehension Network REDINTER: http://www.redinter.eu/web/

5.1. Phraseology in multilingual corpora

The process of searching for the equivalents of phraseological units provides a good illustration of how multilingual corpora can be used in language teaching, dictionary development and translation.

Before we could investigate any phraseological units in the corpus, we need to develop a working definition of a phraseological unit. We decided to work with the definition developed by our colleagues from NASU (The National Academy of Sciences of Ukraine), who are currently working on a Polish-Ukrainian phraseological dictionary. They defined phraseological units as follows:

> Phraseological units are distinguished among other types of phrases by their complicated semantics, which is strongly oriented towards national linguistic worldview. Thus, the main problem in compiling of a bilingual phraseological dictionary is the selection of adequate translational equivalents with due account for differences in worldview represented in the respective language systems. This is why the task of a comprehensive translational phraseological dictionary is to convey the phraseological system of one language by the means of the other language". (Tymoshuk, Vilchynska, Shyrokov and Nadutenko, in press)

As we can see, the only way to provide a description of a phraseological unit is through its ontology, because every language expresses phraseological semantic content in a different manner. Most scholars studying the relations between lexemes and phraseological units argue that a semantic and functional correlation exists between them, which is reflected in the organisation of different levels of language systems. No consensus has yet been attained on how we should determine the position of phraseology among other levels of language systems. V. L. Arkhangel'skyi proposed a structural semantic classification. He defined lexemes and phrasemes as different units organised in a hierarchical relationship, however these units are units of the same level that constitute "building blocks" of sentences [3, pp. 182-188]. This apparent incongruence is a result of the great complexity of the semantics of a phraseological unit and of the priority it takes over a word (after M. M. Shanskyi). A phraseological unit takes the form of a free association of words on the phrase level, whereas on the text level it assumes the role of a word [10, p. 12].

It is equally difficult to clearly delineate the dividing line between a phraseme and a non-phraseme. As a consequence, the selection of phrasemes for contrastive studies is always problematic, because one always needs to decide which linguistic tradition to choose as the source of comparison with other languages. The above applies also to the selection of collocations, which we can also categorise as phrasemes:

> Separated into the so-called "rhombed" zone of the Dictionary are also the collocations – set phrases that allow slight desemantization of one component (eg. **вовчий апетит**), word equivalents (eg. **до безмежжя**) and terminological phrases (eg. **топографічна анатомія**). (Tymoshuk, Vilchynska, Shyrokov and Nadutenko, in press)

Scholars studying phraseology must be prepared to face numerous pitfalls. Idioms that appear strikingly similar may actually carry different, sometimes exactly opposite meanings: Pol. *lekarz z bożej łaski* (= a very bad doctor) / Rus. *милости божьей врач* (= a very good

doctor) [lit. doctor of God's grace], *idzie jak krew z nosa* (= very slowly), *кровь из носу* (= immediately) [lit. flows like blood out of a bleeding nose), *owinąć sobie wokół palca kogoś* (= have somebody under one's command), *обвести вокруг пальца* (= lie to someone in a particularly cunning way) [lit. wrap somebody around one's finger].

The parallel Bulgarian-Polish-Russian corpus allows users to search for phraseological units. We must take into consideration, however, that phrasemes can exist in:

1. Only one language

To ludzie bez ducha, bez dumnych snów, bez wzniosłych porywów. A człowiek bez tego to zwykły tchórz, to szmata.	Те нямат дух, те не знаят какво е горди мечти и горди въжделения, а всеки, който не познава нито едното, нито другото — боже мой! — та той е пълен със страхове и опасения!	У них нет мужества, нет гордости, они не умеют сильно желать. А без этого человек гроша ломаного не стоит.

Bronte, E *Wuthering heights*

Myślałam już nawet, że brak jej piątej klepki. Uciekła do swego pokoju wołając mnie do siebie, chociaż powinnam była ubierać dzieci.	Докато траяха тия неща, по държането й разбрах, че е доста глупавичка. Тя се втурна в стаята си и ме застави да отида при нея, макар че в това време трябваше да обличам децата.	Я приняла ее за полоумную, — так она себя вела, пока совершали обряд: она убежала к себе в комнату и велела мне пойти с нею, хотя мне нужно было переодевать детей.

Bronte, E *Wuthering heights*

2. In two languages

Matki i wychowawczynie - nie żadne lalkowate ślicznotki ze słodkimi ślepkami.	Никакви превзети дамички, никакво въртене на очи!	Только не сентиментальные дамы, не те, что строят глазки.

Wells, H. G. *The War of the Worlds*

W mgnieniu oka wdarłem się na wał i stanąłem na jego koronie. Przede mną leżała twierdza.	След още един миг се бях покатерил по земния насип и стоях на гребена му — вътрешността на редута лежеше в краката ми.	Еще через минуту я взобрался по насыпи и стоял на гребне вала – внутренняя площадка редута была внизу, подо мной.

Wells, H. G. *The War of the Worlds*

3. In three languages

Nigdy nie wyznałem swej miłości słowami, ale jeżeli oczy mają wymowę, to każda gąska musiałaby odgadnąć, że byłem zakochany po uszy.	„Не се признах в любов“2 гласно; и все пак, ако очите могат да говорят, дори един идиот би могъл да долови, че съм влюбен до уши.	Я «не позволял своей любви высказаться вслух»; однако, если взгляды могут говорить, и круглый дурак догадался бы, что я по уши влюблен.

Wells, H. G. *The War of the Worlds*

6. CONCLUSIONS

The data presented above shows how many new insights into phraseology multilingual corpora can provide, even though phraseological units usually exist only in one of the languages being compared. The usual situation is that translators only translate the words in phrasemes. The data also indicates that we need to study equivalents of phrasemes departing from their ontology, following the example of scholars from NASU (see 5.1). It is also worth noting that phrasemes in different languages evoke very different associations and mental images, e.g. Bul. *бързата кучка слепи си ражда* [lit. 'the hasty bitch gives birth to blind pups, Pol. *co nagle to po diable* [lit. 'rush is the devil's thing'], Eng. *haste makes waste.*

During our work on the corpus, we encountered a number of different issues. At the same time, it allowed us to find many new solutions and to introduce some innovations. We have learned that multilingual corpora need to be supplemented with more languages. Every language we add to a corpus enables researchers and practitioners to find new questions relevant to translators, interpreters and language teachers.

References:

DIMITROVA, L., KOSESKA-TOSZEWA, V., 2012. Bulgarian-Polish parallel digital corpus and quantification of time. *Cognitive Studies/Études cognitives*, 12, pp. 199–208.

GARABÍK, R., DIMITROVA, L., KOSESKA-TOSZEWA, V., 2011. Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). *Cognitive Studies/Études cognitives*, 11, pp. 227–239.

GRAMATYKA KONFRONTATYWNA BUŁGARSKO-POLSKA BPCG [GKBP]., 1988–2007. (Vol. 1-12). Sofia, Warsaw.

JASKOT, M. P., 2014. Buscando las brechas de significado: las lagunas léxicas entre el español y el polaco In: Zuzanna Bułat Silva, Monika Głowicka and Justyna Wesoła [eds.] *Variación, contraste,*

circulación. Perspectivas lingüísticas en el hispanismo actual. Acta Universitatis Wratislaviensis. Wrocław, Wydawnictwo Uniwersytetu Wrocławskiego, pp. 127-136.

JASKOT, M. P., (in press). *Lexical non-equivalence in chosen European languages in the context of the policy towards multilingualism in Europe*

KISIEL, A. (in press). Korpusowe badania nad metatekstem. Problem homografii, *Prace Filologiczne*.

KISIEL, A., SATOŁA-STAŚKOWIAK, J., SOSNOWSKI, W., 2014. О работе над многоязычным словарём. *Прикладна лінгвістика та лінгвістичні технології (MEGALING-2013),* pp. 111–121.

KORPUS JĘZYKA BUŁGARSKIEGO IBE BAN, n.d. Available at: http://search.dcl.bas.bg/ [Accessed 6 November 2014].

KORPUS JĘZYKA ROSYJSKIEGO, n.d. Available at: <http://www.ruscorpora.ru/> [Accessed 6 November 2014].

KOSESKA-TOSZEWA, V., 1974. Z problematyki temporalno-aspektowej w języku bułgarskim (Relacja imperfectum - aoryst). *Studia z Filologii Polskiej i Słowiańskiej*, 14, pp. 213–226.

KOSESKA-TOSZEWA, V., 2006. *Gramatyka konfrontatywna bułgarsko-polska* (T. 7: *Semantyczna kategoria czasu*). Warsaw: SOW.

KOSESKA-TOSZEWA, V., GARGOV, G., 1990. *Bylgarsko-polska sypostavitelna gramatika* (T. 2: *Semanticznata kategorija opredelenost/neopredelenost*). Sofia: BAN.

KOSESKA-TOSZEWA, V., MAZURKIEWICZ, A., 1988. Net representation of sentences in natural languages. In: *Advances in Petri Nets. Lecture Notes in Computer Science*, 340*,* pp. 249–266. Berlin: Springer-Verlag.

KOSESKA-TOSZEWA, V., MAZURKIEWICZ, A., 2010. *Time flow and tenses*. Warsaw: SOW.

KOSESKA-TOSZEWA, V. ROSZKO, R., (2015. On Semantic Annotation in CLARIN-PL Parallel Corpora. *Cognitive Studies/Études cognitives*, 15

KOSESKA-TOSZEWA, V., KORYTKOWSKA, M., ROSZKO, R., 2007. *Polsko-bułgarska gramatyka konfrontatywna.* Warsaw: Wydawnictwo Akademickie Dialog.

KOSESKA-TOSZEWA, V., SATOŁA-STAŚKOWIAK, J., SOSNOWSKI, W., 2013. From the problems of dictionaries and multi-lingual corpora. *Cognitive Studies/Études cognitives*, 13, pp. 113–122.

KOSESKA-TOSZEWA, V., SATOŁA-STAŚKOWIAK, J., SOSNOWSKI, W., 2013. О работе над книжными и электронными словарями с польским, болгарским и русским языками. W *Прикладна лінгвістика та лінгвістичні технології (MEGALING-2012)*, pp. 124–135.

ROSZKO, D., 2015. Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej (Na tle literackich języków polskiego i litewskiego). Warsaw: Instytut Slawistyki PAN

ROSZKO, D., 2013. Experimental Corpus of the Lithuanian Local Dialect of Punsk in Poland. Examples of the Lexical and Semantic Annotation. *Cognitive Studies/Études cognitives*, 13, pp. 79–95. DOI: 10.11649/cs.2013.005.

SATOŁA-STAŚKOWIAK, J. 2013, Contemporary Contrastive Studies of Polish, Bulgarian and Russian Neologisms versus Language Corpora, *Cognitive Studies/Études Cognitives,* 13, pp. 143-160.

SATOŁA-STAŚKOWIAK, J. 2014, *Edukacja przyszłych tłumaczy w oparciu o korpusy językowe.* In: Прікладна лінгвістика та лінгвістичні технології, MegaLing-2013: Зб. наук. пр. / НАН України, Укр. мовно-інформ. фонд, Київ, pp.. 211-223.

SATOŁA-STAŚKOWIAK, J., KOSESKA-TOSZEWA, V. 2014, *Współczesny słownik bułgarsko-polski*, Warsaw: Slawistyczny Ośrodek Wydawniczy.

SOSNOWSKI, W., 2013. Forms of address and their meaning in contrast in Polish and Russian languages. *Cognitive Studies/Études cognitives*, 13, pp. 225–235.

TYMOSHUK, R., VILCHYNSKA, K., SHYROKOV, V. and NADUTENKO, M., (in press). *Semantic interpretation of phraseological units in ukrainian-polish electronic phraseological dictionary*

АНТОНОВА, О., ДУБРОВСЬКА, І., and ЛУЧИК, А., 2011. *Українсько-польський словник еквівалентів слова.* (V. Koseska-Toszewa & A. Kisiel, Ed.). Київ: Український комітет славістів, Український мовно-інформаційний фонд НАН України, Національний Університет «Києво-могилянська Академія», Інститут Славістики Польської Академії Наук.

ЛУЧИК, А., АНТОНОВА, О., 2012. *Польсько-український словник еквівалентів слова.* (A. Kisiel & V. Koseska-Toszewa, Ed.). Київ: Український мовно-інформаційний фонд НАН України, Національний Університет «Києво-могилянська Академія», Інститут Славістики Польської Академії Наук.

ШАНСКИЙ Н.М., 1957. Лексика и фразеология современного русского языка: пособие для студентов-заочников факультетов русского языка и литературы педагогических институтов.