http://dx.doi.org/10.18778/2544-1795.02.05

Danuta Roszko

Uniwersytet Warszawski
d.roszko@uw.edu.pl

Roman Roszko

Instytut Slawistyki PAN
roman.roszko@ispan.waw.pl

Wojciech Sosnowski

Instytut Slawistyki PAN
wojciech.sosnowski@ispan.waw.pl

POLSKO-BUŁGARSKIE KORPUSY IS PAN I CLARIN-PL

1. Wstęp

W Instytucie Slawistyki PAN od lat 80. XX wieku prowadzone są badania konfrontatywne z wykorzystaniem semantycznego języka-pośrednika. W początkowym okresie badania były prowadzone w zawężeniu do języków słowiańskich (polskiego, bułgarskiego, serbskiego i chorwackiego). W latach 90. obszar zainteresowań został rozszerzony o języki bałtyckie. Kluczowe prace powstałe z wykorzystaniem metodologii teoretycznych badań konfrontatywnych z językiem-pośrednikiem – to wielotomowa *Gramatyka konfrontatywna bułgarsko-polska* [red. Koseska, Penčev, 1988–2009] oraz *Polsko-bułgarska gramatyka konfrontatywna* [Korytkowska, Koseska-Toszewa, Roszko R. 2007]. Prace fundamentalne, a takimi dziełami są gramatyki, wymagają starannych analiz zasobów językowych. Najlepiej, gdy te zasoby są zgromadzone i opracowane cyfrowo. Pierwsze próby budowy pamięci tłumaczeniowych w Instytucie Slawistyki zostały podjęte przez D. Roszko i R. Roszko. Jeszcze w końcu minionego wieku gromadzili i opracowywali równoległe zasoby dla języków polskiego i litewskiego, które zostały wykorzystane w badaniach języków polskiego i litewskiego [Roszko R. 2004; Roszko D. 2006]. Warto tu podkreślić, że nie tyle zdobyte doświadczenie tych autorów w budowie korpusu polsko-litewskiego, co świadomość korzyści wypływających z zastosowania korpusów w badaniach sprawiły, że kierownik Zespołu Semantyki IS PAN prof. V. Koseska-Toszewa uznała budowę korpusów za zadanie priorytetowe. To doprowadziło do współpracy

Instytutu Slawistyki PAN i Instytut Matematyki i Informatyki BAN. Prof. L. Dimitrova, reprezentująca IMI BAN, wespół z prof. V. Koseską-Toszewą, pokierowały pracami nad Bułgarsko-Polsko-Litewskim Korpusem Tekstów Współczesnych, który wkrótce nazwano Eksperymentalnym Bułgarsko-Polsko-Litewskim Korpusem.

2. Eksperymentalny Bułgarsko-Polsko-Litewski Korpus

Ten Korpus to faktycznie dwa oddzielne korpusy: równoległy i porównawczy. Zespół w składzie L. Dimitrova, V. Koseska-Toszewa, D. Roszko i R. Roszko w latach 2006–2010 opracował zasoby równoległe o objętości przekraczającej 3 500 000 słowoform oraz zasoby porównawcze o objętości 200 000 słowoform. Zespół założył, że wszystkie włączane do korpusu równoległego teksty powinny mieć umocowanie w jednym z trzech języków. Czyli utwór oryginalny powinien powstać w jednym z tych trzech języków a następnie zostać przetłumaczony na pozostałe języki. W trakcie prac okazało się, że to założenie znacznie ogranicza zbiór wspólnych dla tych trzech języków tekstów. Właściwie podstawą do budowy tego korpusu stały się teksty polskie, bowiem tylko one były tłumaczone na pozostałe dwa języki (bułgarski i litewski). Utwory litewskie i bułgarskie są głównie tłumaczone na język polski. W wyniku wyżej opisanych trudności zespół postanowił włączyć do Korpusu teksty powstałe w innym języku.

2.1. Strona techniczna Eksperymentalnego Bułgarsko-Polsko-Litewskiego Korpusu

Pierwszy, podstawowy człon korpusu stanowiły teksty polskie tłumaczone na litewski i bułgarski, które skanowano i konwertowano do formatu TXT. W tym celu posłużono się komercyjnym oprogramowaniem ABBY FineReader. W przypadku przekładów literatury światowej nie zachodziła konieczność skanowania i konwersji. Wszystkie włączone do zasobów Korpusu teksty były anotowane na poziomie paragrafów i zdań, por. Tab. 1, 2.

Tabela 1. Przykład anotacji Eksperymentalnego Bułgarsko-Polsko-Litewskiego Korpusu

<s> Zasobnik zadygotał raz i drugi, zawibrował nieznośnie, drżenie to przeszło przez wszystkie powłoki izolacyjne, przez powietrzne poduszki i wtargnęło w głąb mego ciała – seledynowy kontur wskaźnika rozmazał się. </s>	<s> Контейнерът се разтърси един-два пъти и завибрира непоносимо. Това трептене премина през всички изолирни обвивки, през въздушните възглавници и проникна в тялото ми. Зелените очертания на указателя се размазаха. </s>	<s> Kapsulė suvirpėjo kartą, kitą, paskui ėmė vibruoti, šis nepakenčiamas virpulys perėjo per visas izoliacines plėveles, pripučiamas pagalves ir giliai įsismelkė į mano kūną. Žalsvas indikatoriaus kontūras išskydo. </s>
<s> Patrzyłem na to bez strachu. </s>	<s> Не усещах страх.</s>	<s> Aš nejutau baimės. </s>
<s> Nie przyleciałem z tak daleka, aby zginąć u celu. </s>	<s> Не бях долетял от толкова далеко, за да загина точно пред целта. </s>	<s> Atskridau iš taip toli ne tam, kad žūčiau, pasiekęs tikslą. </s>

Tabela 2. Fragment wynikowego pliku w formacie TMX

```
<tu tuid=”0000000006”>
<tuv xml:lang=”Polish”>
<seg>Wykrzyknik umieszczony w trójkącie ma na celu przedstawienie użytkownikowi ważnych czynności, znajdujących się w instrukcji dołączonej do urządzenia.</seg>
</tuv>
<tuv xml:lang=»Bulgarian»>
<seg>Удивителният знак в равностранен триъгълник има за цел да привлече вниманието на потребителя към важни инструкции за работа и поддръжка (обслужване) в литературата придружаваща устройството.</seg>
</tuv>
<tuv xml:lang=”Lithuanian”>
<seg>
Šauktukas lygiašoniame trikampyje perspėja vartotoją apiesvarbias eksploatavimo ir technièns priežiūros instrukcijas su įrenginiu pateiktoje literatūroje.</seg>
</tuv>
</tu>
<tu tuid=”0000000007”>
<tuv xml:lang=”Polish”>
<seg>Symbol błyskawicy umieszczonej w trójkącie ma na celu ostrzeżenie użytkownika o niebezpiecznym napięciu elektrycznym, które może powodować zagrożenie porażeniem prądem elektrycznym.</seg>
</tuv>
<tuv xml:lang=»Bulgarian»>
<seg>Символът светкавица със стрелка в равностранен триъгълник има за цел да привлече вниманието на потребителя към наличието на не-изолирано високо напрежение вътре в корпуса на продукта, което може да бъде от достатъчна величина за да представлява опасност от електрически урад за потребителя.</seg>
</tuv>
<tuv xml:lang=»Lithuanian»>
<seg>Žaibo ženklas lygiašoniame trikampyje perspėja vartotoją apie neizoliuotą pavojingą įtampą produkto korpuse, kurios pakanka vartotojui sukelti elektros šoką. </seg>
</tuv>
```

Polsko-bułgarski zespół opracował zasady segmentacji tekstów w oparciu o kryterium semantyczne. Dlatego, jak to widać w pierwszym rzędzie w Tab. 1, jednemu polskiemu zdaniu może odpowiadać różna liczba zdań bułgarskich i litewskich. Możliwych wzajemnych wariantów odpowiedniości było oczywiście więcej.

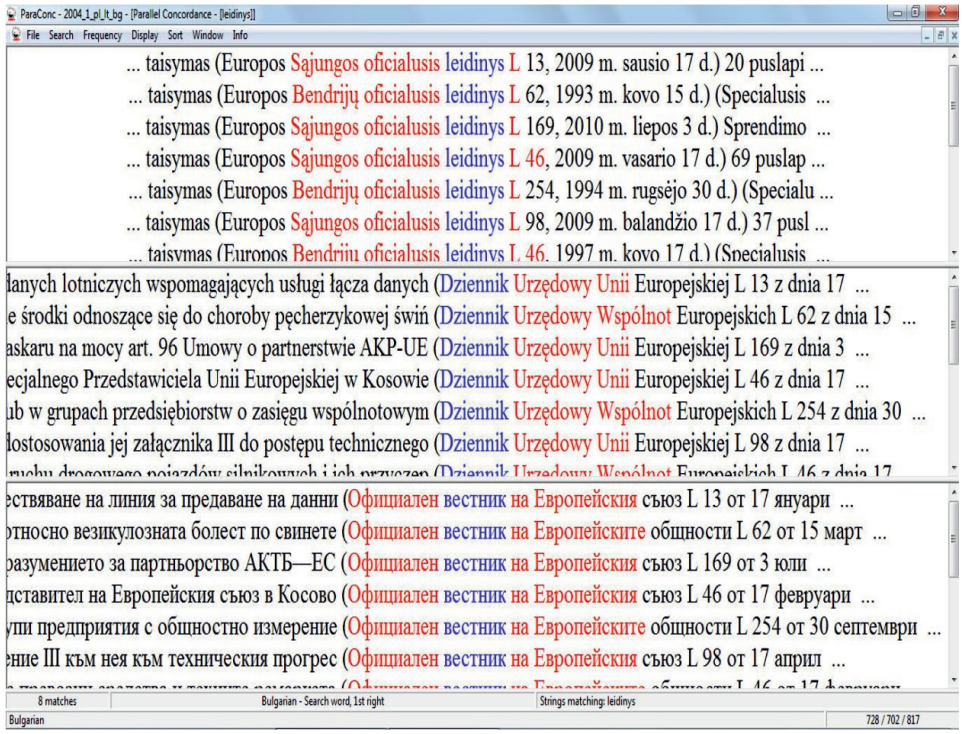
Oprócz segmentacji zespół podjął się otagowania zasobów. W tym celu były wykorzystywano narzędzia TaKIPI (http://nlp.pwr.wroc.pl/narzedzia-i-zasoby/narzedzia/takipi) – dla języka polskiego, MultTex-East (https://www.researchgate.net/publication/266472851_Bulgarian_MULTEXT-East_Corpus_-_Structure_and_Content) – dla języka bułgarskiego oraz MorfoLema (http://donelaitis.vdu.lt/MorfoLema/) – dla języka litewskiego. Docelowo zamierzano wszystkie zasoby opisać w standardzie MULTEXT-East. Próby opracowania znaczników w tym standardzie dla języków polskiego i litewskiego przedstawiono w [Roszko D., Roszko R. 2009; Roszko R. 2009]. Od tego pomysłu odstąpiono. W zamian usta-

lono listę wzajemnych formalnych odpowiedniości między polskim, bułgarskim i litewskim systemem znaczników morfosyntaktycznych.

Równoległa segmentacja tekstów była przeprowadzana jednocześnie w dwóch otwartych oknach narzędzia do zrównoleglania dwujęzycznych zasobów TextAlign (http://mt2007-cat.ru/downloads/TextAlign_1_0_0_3.rar). W jednym oknie opracowywano parę polsko-litewską, w drugim – polsko-bułgarską. Różnice wynikające z odmiennej segmentacji w obu parach wzajemnie uzgadniano. W efekcie otrzymywano dwa wynikowe pliki TMX (polsko-litewski i polsko-bułgarski), które w następnej kolejności konsolidowano, por. Tab. 2.

Roboczo na potrzeby przeszukiwania Eksperymentalnego Bułgarsko-Polsko-Litewskiego Korpusu strona polska korzystała z komercyjnego programu do przeszukiwania wielojęzycznych zasobów ParaConc (http://www.athel.com/para.html), por. Rys. 1.

Rysunek 1. Przykładowy zrzut ekranu ilustrujący wynik wyszukiwania polsko-bułgarsko-litewskich odpowiedników terminologii unijnej



2.2. Znane zastosowania Eksperymentalnego Bułgarsko-Polsko-Litewskiego Korpusu

Zasoby tego trójjęzycznego Korpusu znalazły zastosowanie w badaniach konfrontatywnych polsko-litewskich i polsko-bułgarskich, [por. Dimitrova, Koseska-Toszewa, Roszko D., Roszko R. 2009, 2010, 2014; Koseska-Toszewa, Mazur-

kiewicz 2010; Roszko D. 2015] oraz leksykograficznych polsko-bułgarskich, por. [Satoła-Staśkowiak 2010; Satoła-Staśkowiak, Koseska-Toszewa 2014; Sosnowski, Kisiel, Koseska-Toszewa 2016]. Eksperymentalny Korpus w zamyśle miał być podstawą do budowy trójjęzycznego polsko-bułgarsko-litewskiego słownika on-line. Jednak słownik ten nie ujrzał światła dziennego, choć duża część prac teoretycznych i materiałowych została wykonana, por. Tab. 3.

Tabela 3. Fragment opracowanego słownika polsko-bułgarsko-litewskiego w oparciu o zasoby Eksperymentalnego Bułgarsko-Polsko-Litewskiego Korpusu

<entry> <hw>**сп|я**|</hw> <pos>verb</pos> <gram>imperfect</gram> <conjugation><orth>**-и**|</orth> <type>II</type> </conjugation> <subc>intransitive</subc> <struc type="Sense" n="1"> <trans> **spać** </trans> </struc> <struc type="Derivation" n="1"> <orth>~**и ми се**</orth> <struc type="Sense" n="1"> <trans> **chce mi się spać** </trans> <alt><trans> **ogarnia mnie senność**</trans></alt> </struc> </struc> </entry>	<entry> <hw>**сп|я**|</hw> <pos>verb</pos> <gram>imperfect</gram> <conjugation><orth>**-и**|</orth> <type>II</type> </conjugation> <subc>intransitive</subc> <struc type="Sense" n="1"> <trans>**miegoti**</trans> </struc> <struc type="Derivation" n="1"> <orth>~**и ми се**</orth> <struc type="Sense" n="1"> <trans> **(aš) noriu miego** </trans> <alt><trans> **apima mane miegas** </trans></alt> </struc> </struc> </entry>

3. Baza tekstów współczesnych w językach polskim, bułgarskim i rosyjskim, czyli Polsko-Bułgarsko-Rosyjski Korpus (CLARIN-PL 2013–2016)

Ten Korpus powstawał w Instytucie Slawistyki PAN w ramach zadań realizowanych w projekcie CLARIN-PL w latach 2013–2016. W praktyce zespół IS PAN opracował bazy tłumaczeniowe polskich, bułgarskich i rosyjskich tekstów o łącznej objętości 6 479 367 słowoform. Efekty prac zostały zamieszczone w Repozytorium DSpace CLARIN-PL w formacie TMX (https://clarin-pl.eu/dspace/handle/11321/308) wraz z metadanymi w formacie CMDI. W pracach nad tym Korpusem brał udział zespół IS PAN w składzie: A. Kisiel (do września 2015), V. Koseska-Toszewa, N. Kotsyba, J. Satoła-Staśkowiak, W. Sosnowski.

Udostępnione w repozytorium DSpace zasoby tego Korpusu zawierają oddzielne pliki w formacie TMX dla poszczególnych par: polsko-bułgarskiej i polsko-rosyjskiej. Pierwotnie planowane trójjęzyczne pliki TMX nie cieszyły się zainteresowaniem ze względu na brak popularnych narzędzi do ich obsługi. Ponadto twórcy tego Korpusu zdali sobie sprawę, że liczba użytkowników jednocześnie

zainteresowanych zasobami trójjęzycznymi (polsko-bułgarsko-rosyjskimi) jest zdecydowanie niższa od łącznej liczby użytkowników zasobów dwujęzycznych (polsko-rosyjskich czy polsko-bułgarskich). Konsekwencją przyjętej w założeniach budowy trójjęzycznych zasobów jest uporządkowana jednoczesna segmentacja wszystkich zasobów oraz to, że każdy tekst występuje w trzech wariantach językowych, np. wersja polska: Joseph Conrad, *Lord Jim* (tłumacz Emilia Węsławska), Warszawa 1904 — wersja bułgarska: Конрад Джоузеф, *Лорд Джим* (tłumacz Христо Кънев), София 1968 — wersja rosyjska: Джозеф Конрад, *Лорд Джим* (tłumacz Александра Владимировна Кривцова. Москва 1989).

Wraz z budową wielojęzycznych korpusów V. Koseska-Toszewa i R. Roszko [2015, 2016] opracowali podstawy nowej innowacyjnej anotacji semantycznej zasobów korpusowych z wykorzystaniem jednoznacznych operatorów logicznych, sieci Petriego. W Tab. 4 ukazano przykład anotacji semantycznej.

Tabela 4. Przykład anotacji semantycznej w zdaniach polskim i bułgarskim

Pol.: **< (∃X)P(X)_(state_2)** Był sobie \> **< ?(∃x)(P(x)** krasnoludek \> z długą białą brodą, w czerwonej czapce.
Bułg.: **< (∃X)P(X)_(state_2)** Имаше някога \> **< (∃x) P(x)** едно джудже \> с дълга бяла брадица и алена шапчица.
(∃X)P(X)_(state_2) – kwantyfikacja egzystencjalna stanu
?(∃x)(P(x) – niedopowiedziana kwantyfikacja egzystencjalna obiektu
(∃x) P(x) – kwantyfikacja egzystencjalna obiektu

Polsko-Bułgarsko-Rosyjski Korpus zawiera 162 utwory o zróżnicowanych stylach funkcjonalnych języka. Zauważalna jest przewaga utworów reprezentujących styl artystyczny (beletrystykę) – 78 tekstów oraz urzędowo-kancelaryjny (w tym teksty unijne) – 72 teksty. Pozostałe style mają mniej reprezentantów: styl naukowy – 6 tekstów oraz style prawniczy i religijny – po 3 teksty. Skromna liczba tekstów naukowych, prawniczych i religijnych nie oznacza małej objętości tychże tekstów. 24 teksty to wzajemne tłumaczenia. Pozostałe teksty (138) to tłumaczenia z języków trzecich niereprezentowanych w Korpusie. Segmentacja tekstów, związana z jednoczesnym wyrównaniem tekstów, jest oparta na jednostce znanej z gramatyk szkolnych i akademickich – zdaniu, któremu przypisany jest sens zapewniający zupełność komunikatywną. Ze względu na bliskość typologiczną zestawianych języków nie zachodziła konieczność uszczegółowienia definicji ani opisu natury zdania. W procesie segmentacji tekstów posiłkowano się komercyjnym programem NOVA Text Aligner.

3.1. Znane zastosowania Polsko-Bułgarsko-Rosyjskiego Korpusu (CLARIN-PL 2013–2016)

W tym okresie większość powstających prac na bazie materiałowej tego korpusu dotyczyła zagadnień leksykograficznych, por. [Koseska-Toszewa 2013; Satoła-Staśkowiak, Koseska-Toszewa 2014; Sosnowski, Koseska-Toszewa 2015; Sosnowski, Kisiel, Koseska-Toszewa 2016].

4. Polsko-Bułgarsko-Rosyjsko-Ukraiński Korpus (CLARIN-PL 2016–2018)

Od połowy 2016 roku trwają prace nad rozbudową polsko-bułgarskich zasobów w ramach zadania CLARIN-PL, obejmującego budowę wielojęzycznych korpusów z językiem polskim jako językiem podstawowym. Wynik prowadzonych przez zespół IS PAN w składzie M. Duszkin, D. Roszko, R. Roszko, W. Sosnowski, J. Satoła-Staśkowiak i R. Tymoshuk prac został opublikowany w repozytorium DSpace na stronach CLARIN-PL [Roszko, R., Roszko, D., Sosnowski, Satoła-Staśkowiak 2018] oraz w sieciowej przeglądarce KonText (https://kontext.clarin-pl.eu/run.cgi/first_form).

Ten Korpus jest po części rozwinięciem Korpusu opisanego w pkt. 3. Należy jednak zwrócić uwagę, że została zmieniona koncepcja jego budowy. Obecna struktura nie zakłada istnienia wspólnych dla wszystkich języków zasobów. Takie założenie bardzo ograniczało dobór zasobów. W tym Korpusie segmentacja tekstów nie jest też uzgodniona dla wszystkich języków. Została rozbita na poszczególne pary języków. Językiem dominującym i scalającym całość jest język polski. Poza znacznym zwiększeniem zasobów polskich, bułgarskich i rosyjskich włączono do zasobów korpusowych teksty w języku ukraińskim. Ostateczna planowana objętość Korpusów opracowywanych przez zespół IS PAN w CLARIN-PL przekroczy 50 mln słowoform. Cechą charakterystyczną tego Korpusu (w odniesieniu do wyżej opisanego w pkt. 3.) jest zmiana wewnętrznego zrównoważenia zasobów na korzyść tekstów zawierających terminologię związaną z najnowszymi zdobyczami cywilizacyjnymi i technicznymi oraz potoczną (dialogi filmowe). Zmiany wewnętrznego zrównoważenia zasobów są wynikiem konsultacji oraz wystosowywanych przez obecnych użytkowników baz TMX Korpusu Polsko-Bułgarsko-Rosyjskiego oraz nowych i potencjalnych użytkowników obecnie powstającego Korpusu.

5. CLARIN ERIC i CLARIN-PL

Polsko-bułgarskie zasoby językowe powstają w ramach wielkiego programu badawczego realizowanego pod szyldem CLARIN-PL. CLARIN-PL – to konsorcjum polskich podmiotów naukowych (Instytutu Slawistyki PAN, Instytutu Podstaw Informatyki PAN, Polsko-Japońskiej Akademii Technik Komputerowych, Politechniki Wrocławskiej, Uniwersytetu Łódzkiego i Uniwersytetu Wrocławskiego), współzałożyciel dużej europejskiej infrastruktury naukowej CLARIN ERIC (Common Language Resources & Technology Infrastructure | European Research Infrastructure Consortium). Podstawowym celem CLARIN ERIC jest udostępnianie zasobów i narzędzi językowych dla wszystkich języków europejskich w ramach jednej wspólnej infrastruktury badawczej, stanowiącej warsztat pracy naukowców z nauk społecznych i humanistycznych. Zasoby językowe to bazy danych opisujące w sposób sformalizowany język naturalny w różnych jego aspektach, np. mogą to być korpusy tekstów jedno-, dwu-, wielojęzycznych (dostępne online i przeszukiwalne zbiory tekstów opisane metadanymi lingwistycznymi), słowniki, pamięci

tłumaczeniowe, glosariusze, gramatyki i inne. Natomiast narzędzia językowe to (1) programy do automatycznej analizy tekstu i mowy na różnych poziomach opisu: formalnym (morfologicznym, składniowym), semantycznym i pragmatycznym oraz (2) programy przeznaczone do określonych zadań w przetwarzaniu tekstów (np. do rozpoznawania wystąpień nazw własnych, wyszukiwania terminów i in.).

Nie ulega wątpliwości, że brak zasobów i narzędzi dla określonego języka bardzo ogranicza możliwe zastosowania inżynierii języka naturalnego. Dlatego zespół Instytutu Slawistyki wytrwale zmierza do tworzenia zasobów polskich, bułgarskich, polsko-bułgarskich (i innych) celem umożliwienia analizy tych języków z wykorzystaniem najnowszych systemów przetwarzających język z wykorzystaniem narzędzi dostępnych na stronach CLARIN-PL.

Strategicznym celem infrastruktury CLARIN ERIC jest konsolidacja w jednym sieciowym systemie rozproszonych zasobów, narzędzi językowych oraz usług sieciowych dla wszystkich języków naturalnych stosowanych w Europie. System jest oparty na wspólnych standardach opisu i dostępu oraz udostępniania zebranych (i/lub utworzonych) zasobów i narzędzi językowych naukowcom z obszarów humanistyki i nauk społecznych. W oparciu o potrzeby konkretnych zadań CLARIN projektuje, buduje i udostępnia aplikacje badawcze do pracy ze zbiorami tekstów. Są to działania praktyczne na rzecz rozwoju nowych metod humanistyki cyfrowej i cyfrowych nauk społecznych w wymiarze paneuropejskim, wielojęzycznym i wielokulturowym.

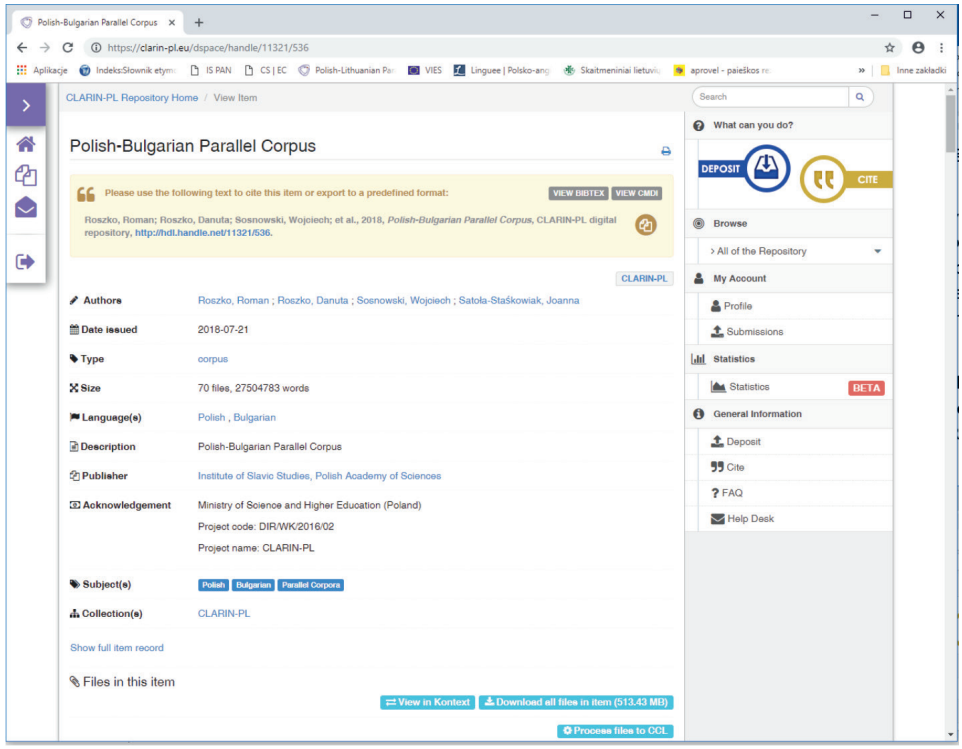
6. Wpływ powstałych korpusów Clarin-PL – IS PAN na funkcjonowanie społeczeństwa i gospodarki

Korpusy równoległe Clarin-PL powstałe w Instytucie Slawistyki PAN (Polsko-Litewski, Polsko-Bułgarski, Polsko-Rosyjski, Polsko-Ukraiński oraz Polsko-Bułgarsko-Rosyjski) w istotny sposób oddziałują na społeczeństwo i gospodarkę. Poza oczywistym zastosowaniem Korpusów Clarin-PL – IS PAN w ośrodkach uniwersyteckich (np. w Polsce są to wyższe szkoły w Warszawie, Poznaniu, Krakowie i Wrocławiu) w celach dydaktycznych, cieszą się one dużym zainteresowanie tłumaczy, pisarzy, redaktorów wydawnictw (najczęściej dwujęzycznych) oraz przedstawicieli mniejszości narodowej.

Na przykład tłumacze przysięgli i przyzakładowi są najbardziej zainteresowani opublikowanymi w wolnym dostępie na stronach Clarin-PL w Repozytorium (DSpace) bazami zrównoleglonych tekstów współczesnych. Jedną z tych baz jest właśnie Polish-Bulgarian Parallel Corpus [Roszko, D., Roszko, R., Sosnowski, Satoła-Staśkowiak 2018], por. rys. 2.

Zaletą wspomnianej bazy jest możliwość zgrania z archiwum DSpace całości zasobów na swój komputer, włączenia do użytkowanych przez siebie programów komputerowo wspomagających proces tłumaczenia (tzw. oprogramowanie CAT) oraz – co ważne – pracy nad ważnymi dokumentami z wyłączonym dostępem do sieci. Jest to niezwykle ważne, zwłaszcza gdy realizowane tłumaczenia zawierają dane wrażliwe, tajne. Zaletą tejże bazy polsko-bułgarskiej (oraz innych opracowy-

Rysunek 2. *Polish-Bulgarian Parallel Corpus* jako zbiór pamięci tłumaczeniowych w formacie TMX jest już dostępny w repozytorium DSpace na stronach Clarin-PL (https://clarin-pl.eu/dspace/handle/11321/536)



wanych przez zespół z IS PAN) jest jej format zapisu danych – najbardziej popularny wśród użytkowników standard pamięci tłumaczeniowej TMX. Uniwersalność tego formatu wynika nie tylko z faktu, że jest najbardziej rozpowszechniony w świecie, lecz również – konwertowalny do innych formatów pamięci TM bez ponoszenia jakichkolwiek kosztów i potrzeby stosowania specjalistycznego oprogramowania. Szczególnie polscy tłumacze zawodowi upodobali sobie program SDL Trados Studio. Pozostali znani nam użytkownicy bazują głównie na systemach CAT oferowanych w wolnym dostępie: memoQ, OmegaT. Wszystkie wymienione tu programy rozpoznają zastosowany przez nas format pamięci tłumaczeniowej TMX.

Część użytkowników (w tym naukowcy i badacze) włączają w/w polsko-bułgarskie zasoby do własnych przeglądarek dwujęzycznych, np. ParaConc.

Polish-Bulgarian Parallel Corpus „2” (Roszko, D., Roszko, R., Sosnowski, Satoła-Staśkowiak 2018) jest dostępny również na stronach Clarin-PL w wielozadaniowej przeglądarce KonText (https://kontext.clarin-pl.eu/run.cgi/first_form, por. rys. 3). W odróżnieniu od w/w baz, które są dostępne dla każdego użytkownika sieci, zasoby korpusowe w środowisku KonText są udostępniane tylko zarejestrowanym użytkownikom Clarin-PL. Ta forma dystrybucji zasobów zapewnia rozbudowane narzędzie równoczesnego przeszukiwania dwujęzycznych.

Rysunek 3. *Polish-Bulgarian Parallel Corpus* (Roszko, D., Roszko, R., Sosnowski, Satoła-Staśkowiak 2018) w środowisku wszechstronnej wyszukiwarki KonText na stronach Clarin-PL (https://kontext.clarin-pl.eu/run.cgi/first_form)

7. Podsumowanie

Korpusy wielojęzyczne mają wiele zastosowań w naukach humanistycznych i społecznych, wpływają też na funkcjonowanie społeczeństwa i gospodarki. Zespół Instytutu Slawistyki PAN stosunkowo wcześnie przystąpił do prac nad korpusami wielojęzycznymi. Należy podkreślić, że ukierunkowanie się w/w zespołu na budowę baz i korpusów nie było ślepym podążaniem za „modą", lecz wynikało bezpośrednio z potrzeb prowadzonych w IS PAN szeroko zakrojonych badań konfrontatywnych języków słowiańskich i bałtyckich. Pierwsze teksty bułgarskie zostały włączone do wielojęzycznych zasobów IS PAN w roku 2006.

Pierwsze korpusy IS PAN miały charakter eksperymentalny. Natomiast budowane od 2013 r. Korpusy CLARIN-PL są w wolnym dostępie.

Zasoby Polsko-Bułgarskiego Korpusu CLARIN-PL są dostępne zarówno w formacie pamięci tłumaczeniowych TMX oraz w przeglądarce on-line KonText.

Bibliografia

Dimitrova L., Koseska-Toszewa V. (2012), *Bulgarian-Polish parallel digital corpus and quantification of time*, „Cognitive Studies | Études cognitives", nr 12, s. 199–208. DOI: https://doi.org/10.11649/cs.2012.013.

Dimitrova L., Koseska-Toszewa V., Roszko D., Roszko R. (2009), *Bulgarian-Polish-Lithuanian Corpus – Current Development*, [w:] C. Vertan, S. Piperidis, E. Paskaleva, M. Slavcheva (Red.), *International Workshop. Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages held in conjunction with The International Conference RANLP-2009, Proceedings*, Borovets, s. 1–8.

Dimitrova L., Koseska-Toszewa V., Roszko D., Roszko R. (2010), *Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus)*, „Cognitive Studies | Études cognitives", nr 10, s. 217–239. DOI: https://dx.doi.org/10.11649/cs.2010.009.

Dimitrova L., Koseska-Toszewa V., Roszko D., Roszko R. (2014). Trilingual Aligned Corpus – Current State and New Applications, „Cognitive Studies | Études cognitives", nr 14, s. 13–20. DOI: https://dx.doi.org/10.11649/cs.2014.002.

Kisiel A., Koseska-Toszewa V., Kotsyba N., Satoła-Staśkowiak J., Sosnowski W. (2016), *Polish-Bulgarian-Russian Parallel Corpus*, CLARIN-PL digital repository, http://hdl.handle.net/11321/308).

Korytkowska M., Koseska-Toszewa V., Roszko R. (2007), *Polsko-bułgarska gramatyka konfrontatywna*, Wydawnictwo Akademickie Dialog, Warszawa.

Koseska-Toszewa V. (2013), *About Certain Semantic Annotation in Parallel Corpora*, „Cognitive Studies | Études cognitives", nr 13, s. 67–78. DOI: https://doi.org/10.11649/cs.2013.004.

Koseska-Toszewa V., Dimitrova L., Roszko R. (Red.) (2009), *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography. MONDILEX Fourth Open Workshop. Warszawa, Poland, 29 June – 1 July, 2009. Proceedings*, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw.

Koseska-Toszewa V., Mazurkiewicz A. (2010), *Constructing catalogue of temporal situations*, „Cognitive Studies | Études cognitives", nr 10, s. 71–109. DOI: https://doi.org/10.11649/cs.2010.004.

Koseska V., Penčev J. (Red.) (1988–2009), *Gramatyka konfrontatywna bułgarsko-polska*, t. I–IX, Sofia–Warszawa.

Koseska-Toszewa V., Roszko R. (2015), *On Semantic Annotation in CLARIN-PL Parallel Corpora*, „Cognitive Studies | Études cognitives", nr 15, s. 211–236. DOI: https://doi.org/10.11649/cs.2015.016.

Koseska-Toszewa V., Roszko R. (2016). *Języki słowiańskie i litewski w korpusach równoległych CLARIN-PL*, „Studia z Filologii Polskiej i Słowiańskiej", nr 51, s. 191–217. DOI: https://doi.org/10.11649/sfps.2016.011.

Koseska-Toszewa V., Satoła-Staśkowiak J., Sosnowski W. (2013), *From the problems of dictionaries and multi-lingual corpora*„Cognitive Studies | Études cognitives", nr 13, s. 113–122. DOI: https://doi.org/10.11649/cs.2013.007.

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*), s. 124–135.

Roszko D. (2006), *Funkcjonalne odpowiedniki litewskiego perfectum w litewskiej gwarze puńskiej i w języku polskim*, Slawistyczny Ośrodek Wydawniczy, Warszawa.

Roszko D. (2015), *Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej (Na tle literackich języków polskiego i litewskiego)*, Instytut Slawistyki PAN, Warszawa.

Roszko D., Roszko R. (2009), *Morphosyntactic Specifications for Polish and Lithuanian. [Description of Morphosyntactic Markers for Polish and Lithuanian Nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)]*, [w:] Koseska-Toszewa V., Dimitrova L., Roszko R. (Red.) (2009), s. 145–158.

Roszko D., Roszko R. (2014), *A net presentation of Lithuanian sentences containing verbal forms with the grammatical suffix -dav-*, „Cognitive Studies | Études cognitives", nr 14, s. 173–182. DOI: https://doi.org/10.11649/cs.2014.014.

Roszko D., Roszko R. (2016), *Polsko-litewskie korpusy równoległe. Elementy anotacji semantycznej z zakresu modalności możliwościowej i kwantyfikacji zakresowej*, [w:] E. Gruszczyńska,

A. Leńko-Szymańska (Red.), *Polskojęzyczne korpusy równoległe. Polish language Parallel Corpora*, Warszawa, s. 119–132. http://repozytorium.ceon.pl/bitstream/handle/123456789/9717/07_Roszko_Roszko.pdf?sequence=1&isAllowed=y, http://rownolegle.blog.ils.uw.edu.pl/files/2016/03/0000_Korpusy.pdf.

Roszko R. (2004). *Semantyczna kategoria określoności/nieokreśloności w języku litewskim (w zestawieniu z językiem polskim)*, Instytut Slawistyki PAN, Warszawa.

Roszko R. (2009), *Description of Morphosyntactic Markers for Polish Verbs within MULTEXTEast Morphosyntactic Specifications (Version 3.0 May 10th, 2004)*, [w:] V. Koseska-Toszewa, L. Dimitrova, R. Roszko (Red.) (2009), s. 157–163.

Roszko D., Roszko R., Sosnowski W., Satoła-Staśkowiak J. (2018), *Polish-Bulgarian Parallel Corpus*, CLARIN-PL digital repository, http://hdl.handle.net/11321/536.

Satoła-Staśkowiak J. (2010), *Polsko-bułgarskie odpowiedniości przekładowe czasów przeszłych*, Instytut Slawistyki PAN, Warszawa.

Satoła-Staśkowiak J. (2013), *Contemporary Contrastive Studies of Polish, Bulgarian and Russian Neologisms versus Language Corpora*, „Cognitive Studies | Études cognitives", nr 13, s. 143–160. DOI: https://doi.org/10.11649/cs.2013.009.

Satoła-Staśkowiak J. (2017), *Badania nad najmłodszą leksyką słowiańską w oparciu o korpusy językowe*, [w:] Д. Благоева, Л. Андрейчин (Red.), *Българско-полски студии*, Българска академия на науките. Институт за български език, s. 32–45.

Satoła-Staśkowiak J., Koseska-Toszewa V. (2014), *Współczesny słownik bułgarsko-polski,* Instytut Slawistyki PAN, Warszawa.

Sosnowski W., 2013. *Forms of address and their meaning in contrast in Polish and Russian languages*, „Cognitive Studies | Études cognitives", nr 13, s. 225–235. DOI: https://doi.org/10.11649/cs.2013.015.

Sosnowski W., Kisiel A., Koseska-Toszewa V. (2017), *Leksykon odpowiedniości semantycznych w języku polskim, bułgarskim i rosyjskim*, Instytut Slawistyki PAN, Warszawa.

Sosnowski W., Koseska-Toszewa V. (2015), *Multilingualism and Dictionaries*, „Cognitive Studies | Études cognitives", nr 15, s. 43–55. DOI: https://doi.org/10.11649/cs.2015.004.

Danuta Roszko, Roman Roszko, Wojciech Sosnowski

POLISH-BULGARIAN CORPORA ISS PAS (IS PAN) AND CLARIN-PL

(Summary)

Multilingual corpora have found many applications in arts and humanities and social sciences, as well as in translation. A number of ways exist in which multilingual corpora can be used. Translators and CAT users would predominantly use translation memories (TM). Other users can choose from two ways of accessing the resources produced by The Institute of Slavic Studies. In the first method, the user needs to download the open-source TMX translation memories from CLARIN-PL DSpace repository (https://clarin-pl.eu/dspace) and load it into their preferred computer application. One can found free and proprietary applications that facilitate querying multilingual corpora; CLARIN-PL also offers free tools. The other method of accessing the multilingual data produced by The Institute of Slavic Studies does not require any advanced computer skills from the user. CLARIN-PL webpage includes the KonText search engine, which contains also Polish-Bulgarian resources (https://kontext.clarin-pl.eu/). The Polish-Bulgarian corpus contains the following types of resources: (1) fiction literature, (2) specialist literature (literature that is a reflection of the latest technological and cultural developments); and (3) film dialogues, which are the most similar to spoken language.

Key words: Polish-Bulgarian Corpora, Parallel Corpora, CLARIN-PL