Danuta Roszko
Uniwersytet Warszawski, Warszawa
E-mail: d.roszko@uw.edu.pl
ORCID: 0000-0001-5566-0522

Roman Roszko
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa
E-mail: roman.roszko@ispan.waw.pl
ORCID: 0000-0002-2291-6939

KORPUSY WIELOJĘZYCZNE WKŁADEM INSTYTUTU SLAWISTYKI POLSKIEJ AKADEMII NAUK W ROZWÓJ INFRASTRUKTURY CLARIN-PL. PRZYKŁADY ANALIZY KORPUSOWEJ NAD WOŁACZEM

1. Wstęp

W Instytucie Slawistyki Polskiej Akademii Nauk[1] (dalej IS PAN) w latach 70.–80. minionego wieku uwaga części językoznawców skierowana została ku badaniom kontrastywnym. Międzynarodowy zespół badawczy, który ze strony polskiej reprezentowali naukowcy IS PAN Kazimierz Feleszko, Małgorzata Korytkowska, Violetta Koseska-Toszewa, Jolanta Mindak, Danuta Rytel-Kuc[2], Irena Sawicka i in., opracował podwaliny nietradycyjnego projektu badań kontrastywnych na potrzeby planowanych wówczas gramatyki konfrontatywnej bułgarsko-polskiej i gramatyki konfrontatywnej serbsko-chorwacko-polskiej (por. *Wstęp*, 1984). Jak twierdzi V. Koseska-Toszewa, był to pionierski projekt konstrukcji semantycznej gramatyki kontrastywnej opartej na logice i semantyce (Косеска-Тошева & Гаргов, 1990). Sami twórcy tę metodę nazwali „teoretycznymi badaniami konfrontatywnymi z semantycznym językiem pośrednikiem",

[1] Podówczas noszący nazwę Instytutu Słowianoznawstwa Polskiej Akademii Nauk.

[2] Obecnie Danuta Rytel-Schwarz (Universität Leipzig).

zaś kluczowe dzieła nosiły tytuły: *Gramatyki konfrontatywnej bułgarsko-polskiej* (Karolak, 2008; Korytkowska & Roszko, 1997; Korytkowska, 1992, 2004; Koseska-Toszewa, 2006; Koseska-Toszewa i in., 1995; Maldjieva, 2009; Maldžieva, 2003; Гугуланова i in., 1993; Косеска-Тошева & *Гаргов*, 1990; Крумова-Цветкова & Рошко, 1994; Петрова-Вашилевич & Чоролеева, 1994; Савицка & Бояджиев, 1988 i in.) i *Polsko-bułgarskiej gramatyki konfrontatywnej* (Koseska-Toszewa i in., 2009). Prowadzenie wyżej wymienionych badań wiązało się ze żmudną ręczną ekscerpcją i analizą, bowiem wówczas nie były dostępne żadne polsko-bułgarskie zasoby korpusowe ani narzędzia językowe. Badacze zdawali sobie sprawę, że cyfrowe dwujęzyczne zasoby są niezbędne w badaniach kontrastywnych. Ta świadomość doprowadziła do pierwszych eksperymentalnych prób zaprzęgnięcia mocy komputerowych do badań kontrastywnych. Piszący te słowa w latach 90. XX wieku przystąpili do budowy namiastki polsko-litewskiego korpusu tekstów współczesnych. Jednak dopiero po roku 2000 prace nad korpusami w IS PAN nabrały właściwego tempa.

2. Początki budowy wielojęzycznych korpusów w IS PAN

Pierwszym dwujęzycznym korpusem był *Eksperymentalny polsko-litewski korpus tekstów dwudziestowiecznych*. Twórcami tego korpusu byli Roman Roszko i Danuta Roszko. Budowa tego korpusu pochłaniała dużo czasu. Konieczne było skanowanie utworów polskich i litewskich, rozpoznanie znaków, korekta rozpoznania oraz zamiana kodowania znaków. Do rozpoznawania znaków stosowano węgierski program Recognita[3], który słabo radził sobie z polskim tekstem i w ogóle nie sprawdzał się z rozpoznaniem tekstu litewskiego. Dlatego wiele godzin poświęcano trenowaniu, tak by tekst polski i tekst litewski mógł być rozpoznany w zadawalającym stopniu. Recognita nie współpracowała z żadnym słownikiem, dlatego rozpoznawanie tekstu całkowicie bazowało na identyfikacji kształtów poszczególnych znaków, liter, ligatur. Jakość druku w Polsce i na Litwie w drugiej połowie XX wieku pozostawiała wiele do życzenia, stąd zdefiniowane wzorce liter mogły być nieskuteczne w dalszych partiach nawet już tej samej książki. Wystarczyła zmiana w ilości nałożonej farby drukarskiej, by program

[3] Recognita – program do optycznego rozpoznawania znaków. W literaturze przedmiotu często stosuje się literowiec OCR (ang. *optical character recognition*) na określenie technik i samego oprogramowania służącego do rozpoznawania tekstu w plikach bitmapowych i jego zapisu do pliku tekstowego. Sam proces rozpoznawania bywa nazywany „ocerowaniem".

tracił wytrenowaną skuteczność „ocerowania". Dużo nakładu pracy wymagała korekta tekstu. Wówczas były dostępne proste narzędzia do sprawdzania polskiej pisowni. Nie było jednak żadnych narzędzi dla języka litewskiego, nie było ani jednego fontu „litewskiego". Ze względu na różnorodność standardów kodowania polskich liter[4] oraz brak jakiegokolwiek standardu kodowania znaków litewskich piszący te słowa postanowili zastosować zapis TeX-owy[5], por. tabela 1. W wierszu pierwszym zamieszczony zostaje zapis z przyjętym standardem TeX-owym, w drugim wierszu – dla porównania – typowy zapis tego tekstu.

Tabela 1. Przykład polsko-litewskiego segmentu w zapisie TeX-owym

	Język polski	**Język litewski**
1.	`Kapelan w kapie sta{\l} u stopni o{\l}tarza i modli{\l} si\k{e} szybko, p{\'o}{\l}g{\l}osem, jakby z obowi\k{a}zku, senny i roztargniony.`	`Kamandorius su kopa ant sav\k{e}s stov\.ejo prie altoriaus ir labai greitai meld\.esi, pus\.en-bals\k{u}, taip kaip i\v{s} prievartos, susn\={u}d\k{e}s ir susimai\v{s}\k{e}s.`
2.	Kapelan w kapie stał u stopni ołtarza i modlił się szybko, półgłosem, jakby z obowiązku, senny i roztargniony.	Kamandorius su kopa ant savęs stovėjo prie altoriaus ir labai greitai meldėsi, pusėn-balsų, taip kaip iš prievartos, susnūdęs ir susimaišęs.

Tak przygotowane zasoby polskie i litewskie były ręcznie zrównoleglane do poziomu zdania w arkuszu kalkulacyjnym Lotus 1-2-3. Tenże program był używany również do gromadzenia i przeszukiwania zasobów. W późniejszym okresie, gdy kodowanie polskich i litewskich znaków zostało unormowane w systemie Windows, zasoby korpusowe zostały dostosowane do standardu wielojęzycznej przeglądarki ParaConc (ParaConc, b.d.) i do niej załadowane. Objętość opisywanego polsko-litewskiego korpusu przekroczyła wielkość 300 000 słowoform. W oparciu o te zasoby piszący te słowa prowadzili własne badania kontrastywne, których efektem były między innymi monografie R. Roszko (R. Roszko, 2004) i D. Roszko (D. Roszko, 2006).

[4] W latach 90. XX wieku nie było wiodącego wzorca kodowania polskich znaków. Stosowano różne standardy, np. ISO 8859-2, IBM (CP852), Mazovia, CSK, Cyfromat, DHN, Logic, IINTE-ISIS, Microvex, IEA-Świerk, Ventura, ELWRO-Junior, Mac, AmigaPL, Atari-Calamus, ATM i inne.

[5] TeX, LaTeX – jest to zestaw różnych makr, znaczników, które składają się na niesamodzielne środowisko programistyczne, służące automatyzacji procesu składu tekstu, por. LaTeX, b.d.

3. Eksperymentalny bułgarsko-polsko-litewski korpus

Eksperymentalny bułgarsko-polsko-litewski korpus powstawał w latach 2006–2010 w dwóch odmianach równoległej i porównawczej. Konstrukcją tego korpusu zajmował się zespół w składzie Ludmiła Dimitrova[6], Violetta Koseska-Toszewa, Danuta Roszko i Roman Roszko[7]. Objętość części korpusu równoległego przekroczyła 3 500 000 słowoform, części porównawczej – 200 000 słowoform. Założenie *Eksperymentalnego bułgarsko-polsko-litewskiego korpusu równoległego* było proste. Wszystkie zamieszczone w korpusie teksty winny powstać po II wojnie światowej w jednym z trzech języków reprezentowanych w korpusie i być przetłumaczone na dwa pozostałe. Ograniczono się do utworów beletrystycznych. Wraz z postępującymi pracami okazało się, że tylko teksty oryginalne polskie i ich tłumaczenia spełniały ten wymóg. Nie stwierdzono bowiem równoległych przekładów z języka litewskiego na polski i bułgarski ani z języka bułgarskiego na polski i litewski. Niestety liczba równoległych tłumaczeń z polskiego na litewski i bułgarski okazała się niewielka, dlatego założenie budowy tego korpusu uległo modyfikacji. Dopuszczono utwory powstałe w języku innym, głównie rosyjskim i angielskim.

Część zasobów tego korpusu była skanowana i „ocerowana" w programie ABBY FineReader. Twórcom tego korpusu zależało na włączeniu konkretnych dzieł, dlatego zdecydowano się na skanowanie i optyczne rozpoznawanie znaków. W przypadku przekładów literatury światowej (rosyjskiej i anglosaskiej) nie zachodziła taka konieczność. Utwory miały dobrą cyfrową reprezentację. Zespół opracował zasady segmentacji tekstów w oparciu o kryterium semantyczne. Wszystkie zasoby były uzgadniane na poziomie paragrafów i zdań[8]. Część zasobów została wzbogacona o anotację morfosyntaktyczną. Dla opisu morfosyntaktycznego języka polskiego zastosowano tager TaKIPI (TaKIPI, b.d. – zgodny z tagsetem Korpusu Instytutu Podstaw Informatyki Polskiej Akademii Nauk), dla języka bułgarskiego – MulTex-East[9] oraz dla języka litewskiego – MorfoLema

[6] Profesor Instytutu Matematyki i Informatyki Bułgarskiej Akademii Nauk. Pomysłodawca i kierownik projektu. Profesor L. Dimitrovą wspomagały jej doktorantki Stefka Kovacheva i Ralitsa Dutsova. Profesor Kiril Simov był konsultantem projektu.

[7] Violetta Koseska-Toszewa, Danuta Roszko i Roman Roszko – reprezentowali IS PAN.

[8] Przykład segmentacji *Bułgarsko-polsko-litewskiego korpusu* oraz fragment opracowanego na jego podstawie *Polsko-bułgarsko-litewskiego słownika* można zobaczyć w D. Roszko i in., 2018.

[9] Były to pierwsze próby opisu morfosyntaktycznego zasobów bułgarskich w tym standardzie. Wybrane zasoby były ręcznie anotowane. Więcej na temat standardu MULTEXT-East dla języka bułgarskiego w Dimitrova i in., 2005, por. też narzędzie do anotacji (Ljubešić i in., 2020). Obecnie nie jest to jedyny anotator zasobów bułgarskich. W Instytucie Języka Bułgarskiego Bułgarskiej Akademii

(MorfoLema, b.d.). Docelowo zamierzano ujednolicić anotację wszystkich zasobów do standardu MULTEXT-East (wstępne opracowanie znaczników w tym standardzie dla języków polskiego i litewskiego: R. Roszko, 2009; D. Roszko & Roszko, 2009). Od tego pomysłu jednak odstąpiono. W zamian ustalono listę wzajemnych formalnych odpowiedniości między dostępnymi tagami polskimi, bułgarskimi i litewskimi.

Ze względu na ograniczenia zastosowanego do segmentacji narzędzia TextAlign (TextAlign, b.d.) twórcy tego korpusu symultanicznie w dwóch otwartych oknach TextAling zrównoleglali zasoby polsko-bułgarskie i polsko-litewskie, uzgadniając wspólną dla trzech języków segmentację. Wyeksportowane pliki wynikowe w formacie TMX[10] (polsko-litewski i polsko-bułgarski) łączono w jeden plik, opisujący zrównoleglenie dla trzech języków.

Ten korpus nie był przewidziany jako twór samodzielny. Miał stanowić punkt wyjścia do budowy pierwszego wielojęzycznego słownika bułgarsko-polsko-litewskiego w wersji on-line. Roboczo na potrzeby przeszukiwania tego korpusu strona polska stosowała ze wspomnianego już wyżej narzędzia ParaConc (więcej informacji na temat tego korpusu, por. Dimitrova i in., 2009a, 2009b, 2010, 2014). Liczne przykłady zastosowania tego korpusu w pracach językoznawczych, por. Duszkin, 2010; Koseska-Toszewa & Mazurkiewicz, 2010; Koseska-Toszewa & Roszko, 2015, 2016; D. Roszko, 2015; Satoła-Staśkowiak, 2010; Satoła-Staśkowiak & Koseska-Toszewa 2014 i in.

4. Wielojęzyczne korpusy równoległe IS PAN – Clarin-PL

4.1. Europejska infrastruktura Clarin ERIC[11]

Dnia 29 września 2006 roku na pierwszej opublikowanej Mapie Drogowej Europejskiej Infrastruktury Badawczej ESFRI[12] znalazła się infrastruktura Clarin, której współzałożycielem było siedem państw, w tym Polska.

Nauk opracowano konkurencyjne do standardu MULTEXT-East narzędzie do tokenizacji, tagowania i lematyzacji zasobów bułgarskich (BgTagger, b.d), por. Koeva & Genov, 2011.

[10] TMX (< ang. *Translation Memory eXchange* / pol. *wymiana pamięci tłumaczeniowej*) – jeden ze standardów zapisu plików wymiany pamięci tłumaczeniowej (TM < ang. *Translation Memory* / pol. *pamięć tłumaczeniowa*).

[11] CLARIN (ang. *Common Language Resources and Technology Infrastructure* / pol. *Wspólne Zasoby Językowe i Infrastruktura Technologiczna*), ERIC (ang. *European Research Infrastructure Consortium* / pol. *Konsorcjum na Rzecz Europejskiej Infrastruktury Badawczej*).

[12] ESFRI (ang. *European Strategy Forum on Research Infrastructures* / pol. *Europejskie Strategiczne Forum na rzecz Infrastruktury Badawczej*).

Obecnie europejską infrastrukturę Clarin tworzy 20 państw i organizacji ponadpaństwowych. Ponadto cztery państwa (Republika Francuska, Republika Islandii, Republika Południowej Afryki i Zjednoczone Królestwo Wielkiej Brytanii i Irlandii Północnej) są obserwatorami. Natomiast Stany Zjednoczone Ameryki uznawane są za część trzecią infrastruktury.

Clarin jest infrastrukturą nowatorską, idealnie wpisującą się w nurt badań interdyscyplinarnych (z pogranicza informatyki technicznej i językoznawstwa), silnie promowanych nie tylko w Europie, lecz również na całym świecie. Infrastruktura Clarin wyrasta z potrzeb użytkowników oraz ogólnoświatowego trendu rozwoju technologii językowych i informatycznych (IT[13]) oraz sztucznej inteligencji (AI[14]). Obszarem AI jest przetwarzanie języka naturalnego, które nie jest możliwe bez ścisłej współpracy językoznawców i informatyków. Ben Gomes (jeden z dyrektorów Google'a) w artykule *Speech recognition is tech's next giant leap, says Google* (Gomes, 2018), zamieszczonym w „The Guardian", wręcz stwierdza, że lingwiści są przyszłością IT.

Strategicznym celem infrastruktury Clarin ERIC jest:

a) konsolidacja w jednym sieciowym systemie rozproszonych zasobów, narzędzi językowych oraz usług sieciowych dla wszystkich języków naturalnych stosowanych w Europie;

b) wytworzenie wspólnych standardów opisu zasobów i narzędzi oraz dostępu do nich;

c) udostępnianie już zebranych oraz powstających zasobów i narzędzi językowych naukowcom z obszarów humanistyki i nauk społecznych.

Na infrastrukturę Clarin ERIC składa się sieć centrów. Są to centra:

- typu A, gdzie powstają podstawy technologiczne i usługi do funkcjonowania sieci;
- typu B, czyli Centrum Technologii Językowych, gdzie użytkownikom dostarczane są narzędzia i zasoby związane z przetwarzaniem języka naturalnego (są to podstawowe elementy sieci);
- typu C, gdzie zawarte są opisy zasobów, czyli metadane w formacie CMDI[15];
- typu K, gdzie użytkownicy otrzymują wsparcie i dostęp do wiedzy oraz ekspertów.

[13] IT (ang. *information technology* / pol. *technologia informatyczna*).

[14] AI (ang. *artificial intelligence* / pol. *sztuczna inteligencja*).

[15] CMDI (ang. *Component Metadata Infrastructure* / pol. *infrastruktura metadanych komponentów*).

4.2. Polskie konsorcjum Clarin-PL

Polska część infrastruktury Clarin, umownie nazywana Clarin-PL, od jej powstania tworzy sieć sześciu instytucji naukowych: Politechnika Wrocławska (lider konsorcjum Clarin-PL), Instytut Podstaw Informatyki Polskiej Akademii Nauk, Instytut Slawistyki Polskiej Akademii Nauk, Polsko-Japońska Akademia Technik Komputerowych, Uniwersytet Łódzki oraz Uniwersytet Wrocławski.

Nadrzędny cel określający budowę polskiej infrastruktury badawczej Clarin-PL to wsparcie rozwoju nauk humanistycznych i społecznych w Polsce w tych obszarach badawczych, których podstawą jest analiza wszelkich (małych i wielkich) danych językowych (pisanych i mówionych). Konsorcjum Clarin-PL tworzy i udostępnia badaczom spójną infrastrukturę, zapewnia wsparcie merytoryczne, dzięki którym jest możliwe prowadzenie badań z wykorzystaniem nowoczesnych metod opartych na technologiach przetwarzania języka (jakościowe oraz ilościowe). Warto podkreślić, że tak prowadzone badania gwarantują badaczom osiąganie wyników mających dostrzegalny wpływ na kształt współczesnej i nowoczesnej nauki światowej.

Pierwsza faza budowy polskiej infrastruktury Clarin-PL przypadła na lata 2013–2018. W tym okresie konsorcjum Clarin-PL trzykrotnie uzyskało wsparcie Ministerstwa Nauki i Szkolnictwa Wyższego. Druga faza rozwoju polskiej infrastruktury Clarin-PL trwa od drugiej połowy 2018 roku. Polega ona na utrzymaniu infrastruktury, ograniczonej jej rozbudowie i przystosowaniu zasobów i narzędzi do zmieniających się standardów światowych. Faza utrzymania jest również finansowana przez Ministerstwo Edukacji i Nauki.

Na początku 2020 roku Konsorcjum Clarin-PL uzyskało dofinansowanie projektu złożonego w Programie Operacyjnym Inteligentny Rozwój 2014–2020, działanie 4.2. Wniosek Konsorcjum Clarin-PL został bardzo wysoko oceniony (druga nota). Kwalifikowany koszt projektu jest bliski 132 milionów złotych. Podstawowym celem tego projektu jest znaczne rozszerzenie skupionej infrastruktury badawczej Clarin-PL, która stanie się platformą badawczo-rozwojową do przetwarzania języka naturalnego i eksploracji wielkich danych językowych (tekstu i mowy) oraz danych multimodalnych.

Rola każdej jednostki naukowej wchodzącej w Konsorcjum Clarin-PL jest znacząca. Na przykład slawiści i bałtyści IS PAN nie tylko budują wielojęzyczne zasoby z językiem polskim jako węzłowym, lecz również współuczestniczą w wypracowaniu koncepcji niezbędnych do modelowania narzędzi językowych, testują te narzędzia oraz weryfikują zasoby.

Wszyscy konsorcjanci promują infrastrukturę Clarin-PL, współorganizują warsztaty (grupowe i indywidualne), na których obecni i potencjalni użytkownicy infrastruktury nie tylko zapoznają się ze stanem obecnym i perspektywami rozwoju

tejże infrastruktury, lecz przede wszystkim zdobywają wiedzę, jak skutecznie korzystać ze wszystkich zgromadzonych w infrastrukturze zasobów i narzędzi.

Brak zasobów i narzędzi dla konkretnego języka bardzo ogranicza możliwe zastosowania inżynierii języka naturalnego. Dlatego językoznawcy IS PAN konsekwentnie uczestniczą w tworzeniu wielojęzycznych zasobów. Powstające zasoby mogą być przez każdego użytkownika infrastruktury Clarin poddane wszechstronnej analizie z wykorzystaniem wszelkich systemów przetwarzających język, wliczając w to opracowane i opublikowane przez polskie konsorcjum Clarin-PL narzędzia.

4.3. Zasoby i narzędzia językowe Clarin-PL

Zasoby językowe to bazy danych opisujące w sposób sformalizowany język naturalny w różnych jego aspektach, np. mogą to być zarówno wielojęzyczne korpusy (ang. *corpora*) i pamięci tłumaczeniowe (TM), jak też słowniki, glosariusze, gramatyki, stochastyczne modele językowe i inne. Wybrane przykłady zasobów językowych, dostępnych na stronie Clarin-PL[16]:

Korpus ChronoPress – portal tekstów prasowych z lat 1940–1962, zawierający około 56 tysięcy starannie dobranych fragmentów tekstów prasowych, opracowanych językowo na poziomie morfosyntaktycznym i ustrukturyzowanych pod względem chronologii.

Korpusy języków słowiańskich i bałtyckich w wyszukiwarce **KonText** – wielojęzyczne korpusy ręcznie anotowane na poziomie zdania i warstwą anotacji fleksyjnej z elementami składniowymi.

Korpus Politechniki Wrocławskiej – polskojęzyczne zasoby o wielowarstwowej anotacji.

Paralela – dwujęzyczny polsko-angielski anotowany korpus równoległy.

Słownik kombinatoryczny Hask – korpus polskich i angielskich tekstów, opisujący związki frazeologiczne.

Słowosieć (plWordNet) – wielki relacyjny słownik semantyczny języka polskiego. Zawiera 191 000 słów, 285 000 znaczeń leksykalno-semantycznych i ponad 600 000 relacji dla języka polskiego. Posiada funkcję słownika polsko-angielskiego (255 000 haseł). Jest największym relacyjnym słownikiem semantycznym w świecie.

[16] http://clarin-pl.eu/ (Clarin-PL, b.d.).

SpokesPL – zbiór (wraz z wyszukiwarką) danych konwersacyjnych zbudowany na bazie 247 580 wypowiedzi liczących łącznie blisko 2,5 miliona słowoform.

Walenty – słownik walencyjny predykatów polskich.

Narzędzia językowe to różne programy on-line (webowe) do automatycznej analizy tekstu i mowy na różnych poziomach opisu: formalnym (morfologicznym, składniowym), semantycznym i pragmatycznym, także przeznaczone do określonych zadań w przetwarzaniu tekstów. Wybrane przykłady narzędzi językowych, dostępnych na stronie Clarin-PL:

Analiza mowy – zestaw narzędzi do analizy mowy polskiej.

Chunker – program do płytkiej analizy składniowej.

ENIAM – kategorialny parser składniowo-semantyczny.

Inforex – system do zarządzanie korpusami tekstowymi w sieci.

Inkluz – narzędzie do wykrywania obcojęzycznych wtrąceń w polskim tekście.

LEM – (Literacki Eksplorator Maszynowy) – narzędzie do przetwarzania tekstów literackich.

MeWeX – narzędzie do wydobywania z korpusów kolokacji wielowyrazowych oraz tworzenia słowników jednostek leksykalnych.

Morfeusz 2 – analizator i generator fleksyjny.

Morpho – bezkontekstowa analiza morfologiczna.

Mowa – narzędzia i usługi do przetwarzania mowy.

NER – narzędzia do rozpoznawania nazw własnych i wyrażeń temporalnych.

Parser zależnościowy – narzędzie analizy składniowej zdań w języku polskim.

POLFIE – zaimplementowana gramatyka LFG języka polskiego.

POLFIE-OT – jak wyżej parser LFG języka polskiego z modułem Optimality Theory zapewniającym automatyczne ujednoznacznienie.

ReSpa – narzędzie do znajdowania słów kluczowych w tekście.

Serel – narzędzie do wyznaczania relacji między nazwami własnymi.

Spatial – narzędzie do rozpoznawania relacji przestrzennych w polskich tekstach.

Summarize – narzędzie do streszczania tekstów.

Tagger WCRFT2 – tokenizacja i tagowanie morfosyntaktyczne.

TermoPL – narzędzie do wykrywania obcojęzycznych wtrąceń w tekście.

WebSty i **WebSim** – wielojęzyczne systemy do analizy stylometrycznej, statystycznej analizy semantycznej tekstów oraz analizy podobieństwa tekstów.

WiKNN – (= Wikipedia K-Nearest Neighbours) klasyfikator tematyczny tekstów polskich i angielskich.

WNLoom-Viewer – aplikacja desktopowa do przeglądania Słowosieci.

WoSeDon – narzędzie do ujednoznaczniania znaczeń leksykalnych w tekście poprzez odniesienie do Słowosieci.

WSD – narzędzie do ujednoznaczniania znaczeń leksykalnych.

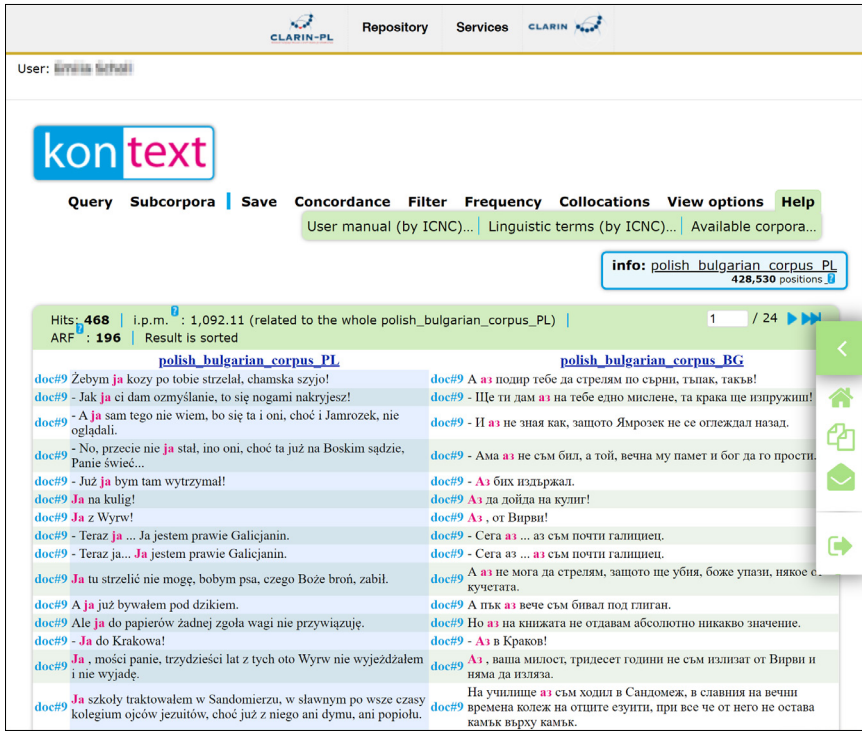
4.4. Wielojęzyczne korpusy z centralnym językiem polskim Clarin-PL

W ramach prac na rzecz infrastruktury Clarin-PL zespół IS PAN opracowuje zasoby dla badaczy szeroko pojmowanych nauk humanistycznych i społecznych, także wykładowców uniwersyteckich i tłumaczy przysięgłych. Są to dwu- i trójjęzyczne korpusy języków słowiańskich i bałtyckich. Korpusy dwujęzyczne o objętości ponad 50 milionów słowoform łączy język polski. Do 2018 roku opracowano dwujęzyczne korpusy tekstów równoległych: polsko-litewski (16 543 470 słowoform), polsko-bułgarski (27 504 783), polsko-rosyjski (5 615 274) i polsko-ukraiński (1 156 579). Te korpusy są dostępne w Repozytorium Clarin-PL dSpace w formacie TMX wraz z metadanymi CMDI pod adresami: https://clarin-pl.eu/dspace/handle/11321/536 (*Polish-Bulgarian Parallel Corpus*), https://clarin-pl.eu/dspace/handle/11321/539 (*Polish-Lithuanian Parallel Corpus* „2"), https://clarin-pl.eu/dspace/handle/11321/534 (*Polish-Russian Parallel Corpus*), https://clarin-pl.eu/dspace/handle/11321/535 (*Polish-Ukrainian Parallel Corpus*) oraz na stronie Clarin-PL w wielojęzycznej przeglądarce KonText pod adresem https://kontext.clarin-pl.eu/run.cgi/first_form. Aby uzyskać dostęp do zasobów w przeglądarce KonText, wymagana jest rejestracja użytkownika na stronie Clarin-PL (https://ctj.clarin-pl.eu/auth/). Na rysunku 1 przedstawiono przykładowy wynik wyszukiwania równoległego w *Polish-Bulgarian Parallel Corpus*:

> polskiego *ja* [o anotacji `ppron12:sg:nom:m1:pri`[17]] i bułgarskiego *аз* [o anotacji `P аз PHYs1`[18]].

[17] Zasoby polskie wszystkich korpusów są otagowane Tagger-em WCRFT2 (Tagger WCRFT2, b.d.), narzędziem rozwijanym przez konsorcjum Clarin-PL.

[18] Zasoby bułgarskie *Polish-Bulgarian Parallel Corpus* początkowo były tagowane BgTagger-em (http://dcl.bas.bg/dclservices/index.php), opracowanym w Instytucie Języka Bułgarskiego Bułgarskiej Akademii Nauk. W grudniu 2020 roku zapadła decyzja o zmianie narzędzia BgTagger na BTB-Pipe (b.d.).



Rys. 1. *Polish-Bulgarian Parallel Corpus* w przeglądarce webowej KonText. Wynik (widoczny fragment strony 1 z 24) został uzyskany w odpowiedzi na zadane polecenie jednoczesnego wyszukania przykładów użycia zaimków polskiego *ja* i bułgarskiego *аз*, obu w mianowniku liczby pojedynczej

Interfejs programu webowego KonText jest skalowalny, automatycznie dopasowuje się do urządzenia (smartfon, tablet, laptop) oraz wielkości i rozdzielczości okna. Umieszczone w górnej części okna poziome Menu jest przejrzyste. Po wybraniu dowolnej z dziewięciu pozycji Menu poniżej pojawia się linia Podmenu, skojarzona z dokonanym wyborem, por. na rysunku 1 wyróżniony zielonym podświetleniem element „Help" wraz z trzema do wyboru opcjami w linii poniżej. Wybór podopcji powoduje wyświetlenie okna, na którym użytkownik może precyzować swoje preferencje (np. wyszukiwania, wyświetlania, sortowania, zapisu wyników w formacie CSV, XML lub TXT, zawężania zasobów do przeszukania, filtrowania i uszczegóławiania warunków już uzyskanej odpowiedzi itd.). Cechą narzędzia KonText jest możliwość analizy również zasobów tylko jednego języka.

Szerzej o przeglądarce KonText można przeczytać w materiałach pochodzących z warsztatów Clarin-PL, które odbyły się w Lublinie w roku 2019 (R. Roszko, 2019) oraz obejrzeć na ten temat tutorial na YouTube z warsztatów Clarin-PL z roku 2020 (R. Roszko, 2020).

Korpusy z centralnym językiem polskim Clarin-PL są rozwijane. Dodawane są nowe teksty. Rozbudowie podlega niewidoczne dla użytkownika acz istotne dla funkcjonalności korpusów znakowanie poszczególnych słowoform tychże zasobów. Jest tu mowa o tzw. lematyzacji[19] i tagowaniu[20].

Celem lematyzacji i tagowania jest umieszczenie oznaczeń, które stają się referencją do zdefiniowanych bloków danych. Wprowadzone do korpusu oznaczanie pozwala użytkownikowi precyzować pytania w przeglądarce KonText. Poniżej zostanie opisany konkretny przypadek. Użytkownik polsko-litewskiego korpusu zamierza badać litewskie deminutywa, będące nazwami pospolitymi i zawierające sufiks *-el-*. W najprostszych przeglądarkach (np. Linguee, b.d.) jest to zadanie niemożliwe do zrealizowania. Ów program wyszukuje tylko według słów i fraz. Od użytkownika wymagane jest podanie realnie istniejącej formy, np. *namelis* 'domek'. W odpowiedzi na zapytanie *namelis* 'domek' pojawią się wszelkie słowoformy leksemu *namelis* 'domek'. Akurat Linguee.com nie wymaga podania formy bazowej, by wszelkie formy odmienne konkretnego leksemu zostały wyświetlone. Zatem po wpisaniu w oknie wyszukiwania słowoformy *namelių* ('domków', gen. pl.) uzyskamy zbliżone wyniki do zapytania *namelis* ('domek', nom. sg.). Różnice dostrzeżemy w kolejności wyświetlanych przykładów, w przypadku zapytania *namelis* w pierwszej kolejności pojawią się zdania zawierające słowoformę *namelis*, w przypadku *namelių* zaś – te zawierające słowoformę *namelių*.

Przeglądarki, w których zaimplementowano stosowanie wyrażeń regularnych, udostępniają użytkownikowi więcej możliwości, por. korpusy InterCorp (https://kontext.korpus.cz/corpora/corplist?requestable=1). Na przykład, wpisanie w oknie wyszukiwania takiej sekwencji

i. ".*el(is|io|iui|į|iu|yje|yj|iai|ių|iams|ius|iais|iuose|iuos|ė|ės|ei|ę|e|ėj|ėje|ių|ėms|es|ėmis|ėse)",

[19] Lematyzacja polega na przypisaniu każdej słowoformie słowa bazowego (lematu). Na przykład, słowoforma *samochodów* zostaje opisana lematem [samochód].

[20] Tagowanie polega na oznaczeniu cech morfosyntaktycznych właściwych słowoformie. Na przykład, słowoforma *samochodów* zostaje opisana zestawem znaczników [subst:pl:gen:m3], co oznacza, że jest to [rzeczownik : w liczbie mnogiej : w dopełniaczu : rodzaju męskiego nieożywionego/przedmiotowego].

pozwala użytkownikowi uzyskać wykaz wszelkich form, które spełniają następujące warunki: [forma złożona z dowolnej sekwencji znaków (również zerowa) poprzedzająca „el"] + [tylko teoretycznie sufiks „el"] + [następująca po „el" jedna z wymienionych w nawiasie sekwencji znaków]. Wynik takiego zapytania może zawierać leksemy: *Briuselis* 'Bruksela', *Izraelis* 'Izrael', *kelias* 'droga', *kelis* 'kilka', *didelis* 'duży', *butelis* 'butelka', *langelis* 'okienko', *vamzdelis* 'rurka' i in. Rozdzielone pionowymi kreskami alternatywne sekwencje znaków – to wszystkie możliwe postaci fleksji rzeczownikowej dla paradygmatu *-elis*, *-elė*.

Aby uzyskać identyczny wynik w przeglądarce KonText[21] Clarin-PL, wystarczy w oknie wyszukiwania wpisać:

ii. `".*el(is|ė)"`

(gdy domyślnie ustawione jest wyszukiwanie według leksemów) lub po prostu

iii. `[lexem=".*el(is|ė)"]`

Porównanie złożoności składni zapytania (i) z (ii) / (iii) wypada na korzyść zastosowanej w Clarin-PL przeglądarki KonText. To skrócenie zapytania stało się możliwe dzięki zastosowanej lematyzacji zasobów korpusowych.

Powróćmy do form podanych w charakterze przykładowych wyników zapytania. Można zauważyć, że takie formy, jak *Briuselis* 'Bruksela', *Izraelis* 'Izrael', są, po pierwsze, nazwami własnymi, po drugie – nie zawierają sufiksu deminutywnego *-el-*. Podobnie w następujących przykładach *kelias* 'droga', *kelis* 'kilka' i *butelis* 'butelka' wyróżnione *-el-* nie jest sufiksem, lecz częścią tematu/rdzenia. Ponadto *kelis* 'kilka', podobnie jak *didelis* 'duży' – nie są rzeczownikami. Praktycznie tylko dwa ostatnie leksemy *langelis* 'okienko' i *vamzdelis* 'rurka' są leksemami, które spełniają oczekiwania użytkownika, tj. wyszukiwana forma jest deminutywnym rzeczownikiem pospolitym z sufiksem *-el-*. W wyszukiwarce KonText możemy doprecyzować zapytanie, by wyświetlone wyniki bardziej odpowiadały oczekiwaniom odbiorcy. By zawęzić wyświetlanie słowoform do rzeczowników, należy dodać tag o wartości rzeczownik. W nomenklaturze litewskiej stosowany jest skrót dkt (< *daiktavardis* 'rzeczownik') na oznaczenie rzeczowników, dlatego prawidłowa forma zapisu przybiera postać tag="dkt.*". Następnie należy ograniczyć wyniki do nazw pospolitych. Litewski tagger oznacza tylko nazwy własne, dlatego

[21] Zasoby czeskiego InterCorp (https://kontext.korpus.cz/corpora/corplist?requestable=1) również bazują na przeglądarce KonText (Machálek, 2020). Nie wszystkie jednak zasoby korpusowe InterCorp zostały przystosowane do wyszukiwania według lematów i tagów, np. zasoby litewskie nie zostały otagowane, ale polskie i bułgarskie – tak. InterCorp jest częścią *Narodowego Korpusu Języka Czeskiego* (Cvrček & Richterová, 2020).

użytkownik powinien skorzystać ze składni negacji, by w odpowiedzi uzyskać tylko nazwy pospolite. Właściwa postać takiego zapisu to tag!="tkr.*", gdzie skrót tkr (< *tikrinis* 'własny') jest oznaczeniem nazwy własnej. O negacji nazwy własnej świadczy „!" w formule tag!.

Łącząc wszystkie wymagane składniki spójnikiem & 'i', uzyskuje się następującą składnię zapytania:

iv. `[lexem=".*el(is|ė)" & tag="dkt.*" & tag!="tkr.*"]`

Włączenie do składni zapytań znaczników (tagów, por. iv) jest możliwe dzięki przeprowadzonej anotacji zasobów korpusów Clarin-PL. Tu jednak należy wyjaśnić, że zakres anotacji słowoform w tych korpusach póki co jest ograniczony do parametrów morfosyntaktycznych. Parametry, takie jak składniowe, semantyczne nie zostały uwzględnione.

4.5. Realizowane i planowane prace

Zespół IS PAN nieustannie pracuje nad rozbudową już powstałych korpusów. Jednocześnie usuwa dostrzeżone błędy w pisowni, zrównolegleniu i anotacji zasobów. Rozszerza funkcje przeglądarki KonText.

W 2019 roku Zespół IS PAN przystąpił do budowy nowych korpusów: litewsko-bułgarskiego, litewsko-rosyjskiego, litewsko-ukraińskiego, bułgarsko-rosyjskiego, bułgarsko-ukraińskiego oraz rosyjsko-ukraińskiego.

W drugiej połowie 2020 roku podjęte zostały prace nad nowymi quasi-referencyjnymi korpusami równoległymi: polsko-bułgarskim, polsko-litewskim, polsko-rosyjskim i polsko-słoweńskim. Zasoby do tych korpusów zostaną dobrane z dbałością o wewnętrzne zrównoważenie. Zrównoleglenie oraz anotacja zostaną wprowadzone ręcznie przez trójosobowe zespoły (dwóch specjalistów niezależnie od siebie anotuje zasoby, trzeci zaś – sprawdza zgodność obu anotacji; w przypadkach rozbieżności – wskazuje właściwy opis).

4.6. Zastosowania korpusów wielojęzycznych IS PAN – Clarin-PL

Potencjalne zastosowania wielojęzycznych korpusów zostały zwięźle przedstawione w podpunkcie 4.4. W tej części artykułu zostaną przytoczone niektóre ze znanych nam zastosowań wielojęzycznych korpusów Clarin-PL. Po opublikowaniu korpusów w Repozytorium Clarin-PL pierwszymi stałymi użytkownikami zostali tłumacze przysięgli i przyzakładowi z Polski i Litwy. Udostępnione przez konsorcjum Clarin-PL pamięci tłumaczeniowe (TM) z zasobami korpusowymi

zostały dostosowane do posiadanego przez tłumaczy oprogramowania CAT[22] i zainstalowane na jednostkach lokalnych. Kolejnymi stałymi zarejestrowanymi użytkownikami korpusów Clarin-PL są lektorzy i wykładowcy uniwersyteccy (z ośrodków naukowych w Krakowie, Poznaniu, Słupsku, Warszawie, Wrocławiu oraz kilku na Ukrainie i Litwie) a także nauczyciele szkół podstawowych i średnich (z Puńska – korpus polsko-litewski i Krakowa – korpus polsko-ukraiński). W oparciu o cytowania wiadomo, że korpusy Clarin-PL znajdują zastosowanie w badaniach kontrastywnych, korpusowych, leksykalnych i leksykograficznych, prowadzonych przez badaczy w Europie i Azji.

5. Ilustracja potencjalnego zastosowania korpusów wielojęzycznych IS PAN – Clarin-PL w analizie bułgarsko-polsko-litewskiej

Не съм достатъчно стар, за да свидетелствам за времето, от което дори най-образованите ни сънародници – съзнателно или не – се заемат с умъртвяването на звателния падеж в българския език. Стаменов, И. (2018, czerwiec 27). Звателният падеж и новият раздел „Език свещен…". От Извора. https://www.otizvora.com/2018/06/9965/

W literaturze przedmiotu zagadnienie wołacza wywołuje ożywione dyskusje. Ostatnio szerzej na temat istoty wołacza pisał między innymi Jan Skarbek-Kazanecki (Skarbek-Kazanecki, 2016). Upraszczając sprawę, z jednej strony uważa się, że wołacz jest jednym z wielu przypadków (pierwotne założenie), z drugiej zaś – że nim nie jest. Jednym z badaczy, który opowiadał się za wykreśleniem wołacza z listy przypadków był wybitny indoeuropeista Jerzy Kuryłowicz (Kuryłowicz, 1949, 1968). Uwzględnienie zatem w kontekście wołacza analizy użycia jego form w takich językach, jak litewski, polski czy bułgarski, może być wartością dodaną w rozstrzyganiu teoretycznych sporów nad tą kategorią. W wielu pracach indoeuropeistycznych wołacz zaliczany jest do systemu deklinacyjnego języka praindoeuropejskiego. Ponadto powszechnie uważa się, że litewskie formy imienne są bardzo zachowawcze. Dlatego badacz może spodziewać się w tym języku dobrze zachowanego wołacza. Z kolei język polski, zwłaszcza w odniesieniu do litewskiego, sprawia wrażenie języka o bardziej wyeksponowanych cechach analitycznych. To potencjalnie mogło prowadzić do zaniku użycia wołacza. Na koniec, język buł-

[22] CAT (< ang. *computer-assisted translation* / pol. *tłumaczenie wspomagane komputerowo*).

garski uważa się za język analityczny[23], który teoretycznie wraz z uproszczeniem i późniejszą utratą deklinacji imiennej nie powinien zachować wołacza, a jednak go utrzymał, por. bułg. *господине*, *брате*, *Иване*, *Боже*, *сине*, *човече*, *отче*, *бабо*, *майко*, *лельо*, *учителю* i in. O zaniku systemu deklinacyjnego rzeczownika świadczą skostniałe formy dawnych przypadków: bułg. *майце* [< *майка*], *тем* [< *те*] *подобни*, *мене*, *днес*, *снощи*, *преди*/*след Христа*, *долу*, *утре* i in. (por. Бояджиев i in., 1983).

Do analizy wybrano przykłady[24] z *Eksperymentalnego bułgarsko-polsko-litewskiego korpusu*:

[1][25] lt – Pasiruošęs, *Kelvin**ai***? – pasigirdo ausinėse.
– Pasiruošęs, *Modard**ai***, – atsakiau.
PL[26] – Gotów, *Kelvin*? – rozległo się w słuchawkach.
– Gotów, *Moddard* – odpowiedziałem.
bg – Готов ли си, *Келвин*? – разнесе се в слушалките.
– Готов съм, *Модард* – отговорих.

[2] lt – *Snaut**ai***… – sukuždėjau.
PL – *Snaut*… – szepnąłem.
bg – *Снаут*… – прошепнах аз.

[3] lt – Argi tai svarbu, *Kris**ai***?
PL – Czy to ważne, *Kris*?
bg – *Има* ли значение това, *Крис*?

[4] lt – *Har**e***…?
PL – *Harey*…?
bg – *Харей*?…

We wszystkich przykładach [1–4] pochyleniem zaznaczono nazwy własne, które potencjalnie w analizowanych tu językach mogłyby wystąpić w wołaczu. Nietrudno nie zauważyć, że tylko formy litewskie są w wołaczu (por. voc. *Kelvin-**ai*** a nom. *Kelvin-as*). W podanych przykładach wykładnik morfologiczny wołacza

[23] Abstrahujemy od ścisłych warunków uznania języka za analityczny.

[24] Tytuł oryginału Stanisław Lem, *Solaris* (1961). Tłumaczenia: na język bułgarski – Станислав Лем, *Соларис* (1965, tłumaczka Andreana Radeva), na język litewski – Stanislavas Lemas, *Soliaris* (1978, tłumaczka Giedrė Juodvalkytė).

[25] Gdy zasoby korpusowe są stosowane tylko do ilustracji opisywanego zagadnienia, wówczas mówi się o badaniach, w których korpusy są jedynie źródłem przykładów. W anglojęzycznej literaturze określa się to terminem ang. *corpus-illustrated approach* / *corpus-informed approach*.

[26] Wielkie litery identyfikatora języka wskazują na język oryginału. Małe litery identyfikatora wskazują na tłumaczenie.

litewskiego został wytłuszczony. Polskie i bułgarskie imiona występują w postaci mianownikowej, jednak oddzielenie tych form przecinkami sugeruje, że możemy mieć do czynienia z wołaczem równym mianownikowi.

Należy podkreślić, że żadne imię/nazwisko ani w języku bułgarskim, ani w języku polskim w tym utworze nie zostało użyte w formie wołacza. Tylko w tłumaczeniu litewskim konsekwentnie stosowane są formy wołacza. Poniżej inne wyekscerpowane z tegoż utworu przykłady, w których przynajmniej w jednym języku można mówić o użyciu formy wołacza:

[5]

lt. *O dangau*	– pl. *wielkie nieba*	– bg. *Боже мой*
lt. *O viešpatie!*	– pl. *Wielki Boże!*	– bg. *Боже мой!*
lt. *Dievuliau!*	– pl. *dobry Boże!*	– bg. *Боже мой!*
lt. *žmogau*	– pl. *człowieku*	– bg. ø
lt. *Ką, mielasis*?	– pl. *Co, miły?*	– bg. *Какво, мили*[27]*?*
lt. *mieloji*	– pl. *kochanie*	– bg. *мила*
lt. *mielas berneli*	– pl. *kochany chłopcze*	– bg. *мило момче*
lt. *vaikuti*	– pl. *mój mały*	– bg. *мое малко момче*
lt. *vaikut*	– pl. *mój maleńki*	– bg. *мъничък мой*
lt. *nepalaužiamas nugalėtojau*	– pl. *niezłomny zdobywco*	– bg. *упорит завоевателю*

Niektóre litewskie formy wołacza są utworzone od deminutywów (np. litew. *Diev-**ul**-iau*, *vaik-**ut**-i*), inne – to formy imienne złożone (np. *mielas-**is***, *mielo-**ji***). Ogólna liczba użyć litewskiego wołacza w tym utworze trzykrotnie przewyższa łączną liczbę użyć form wołacza w językach polskim i bułgarskim razem wziętych.

Analiza kontrastywna innych, wcześniejszych dzieł polskich i ich tłumaczeń na język litewski, dostarcza nowych faktów. Formy wołacza zazwyczaj regularnie pojawiają się w obu językach[28]:

[6] lt – *Broli Bernardai*! – atsiliepė kaip užpykęs ir rodos paskutinį žodį ištarė…
pl – *Bracie Bernardzie*! – odezwał się, wybuchając, jak zmuszony żywym temperamentem…

[7] lt – Reikia – sako – *mano kūdiki*, melstis.
pl – Potrzeba – odezwał się – *dziecko moje*, modlić się.

[27] Współcześnie w języku bułgarskim rejestruje się formy przymiotnikowe w wołaczu, tak jak *мили*, por. bułg. *мили синко* i in. Są to dawne długie/pełne formy z fleksją *-и*.

[28] Józef Kraszewski, *Kunigas* (1882). Juozas Ignotas Kraševskis, *Kunigas* (1887, tłumacz Augustinas Zeicas).

[8] lt – Gerai, *tėveli*, į kelionę tai kelionę! – tarė ne labai gera vokiška kalba Šventas, kuris nuolatos šypsojosi, rodydamas aštrius dantis.
pl – Dobrze, *ojczulku*, w drogę to i w drogę! – z akcentem jakimś obcym, po niemiecku, łamaną mową, począł schrypły Szwentas, który śmiał się ciągle i długi rząd małych, ostrych zębów pokazywał.

[9] lt – Nes žiūrėk! žiūrėk, – kalbėjo Bernardas, – tu *gyvuli lietuviškas*, idant tau nepasidabotų tų laukinių būdas, kad manęs ir zokono neprigautum!
pl – Ale patrz! patrz – mówił Bernard – ty *bestio litewska*, aby ci, na wolność puszczonemu, nie zasmakowało dzikie życie i dawny sprośny obyczaj, abyś mnie nie zdradził i Zakonu!

Są też rejestrowane przykłady przedstawiające dodanie w tłumaczeniu na język litewski leksemu w wołaczu, któremu formalnie nie odpowiada żadna forma w polskiej wersji utworu:

[10] lt Ak, *broli*, žmonės pamena, kaip dainuodavo, puikiai rėdydavos, o kūdikį kaišydavo į kvietukas, supdavo lopšyje.
pl Pamiętają ludzie, gdy jeno pieśni nuciła, strojno chodziła, a chłopiątko w kwiatki ubierając, na ręku kołysała.

[11] lt O čia, *vyreli*, Svalgūną paleido!
pl A jeszcze poznawszy Swalgona…

[12] lt Ak, *dievaiti dieve mano*, Kur aš rasiu avį savo.
pl A! któż mi szukać pomoże owieczki mojej jedynej?

Spójrzmy na inne przykłady[29]:

[13] lt Kad jis pasakys tau: „*Jonai*, duok man už tą bylą šimtą rublių", tai reiškia […]
pl Kiedy on tobie powie: „***Jasiuk***, daj mnie na ten interes sto rubli", znaczy […]

[14] lt – Ai, *Joneli*, *Joneli*!
pl – Oj ***Jasiuk***, ***Jasiuk***!

[15] lt – Negaliu eiti namo, *Joneli*, negaliu namo! – tęsė moteriškė.
pl – Nie mogę ja do chaty iść, ***Jaśku***, nie mogę do chaty! – zawiodła kobieta.

[16] lt Oi, *Pranukai*, *Pranukai*! Savo palaidojimui laikiau aš tuos pinigus, ne pavargėlės palaidojimui, krikščioniškam ir gražiam kapui, kad uždengtų mano gėdą, kurią kentėjau per visą gyvenimą…
pl Oj *Pilipku*, *Pilipku*! na śmierć ja sobie te hrosze chowała, na śmierć dostatnią, chrześcijańską i mogiłę śliczną, coby mi nagrodziła wstyd, który żyjąca piłam…

[29] Eliza Orzeszkowa, *Niziny* (1885), Eliza Ožeškienė, *Šunadvokatis* (1901, tłumacze Vincas Kalnietis i Jonas Jablonskis).

[17] lt – Eik tu, *beproti*! – sušuko svečias – tu jau manai, kad tai ir pasibaigė, prapuolė.
pl – Oj ty, *durniu*! – z dziwną raźnością i energią zawołał sołdat – i ty sobie myślisz, że to już skończone, zapieczętowanie i przepadło?

[18] lt – Klausyk, *Mikalojau*, – rečiau, kaip pirma, pradėjo kalbėti – ar labai brangus tas jūsų advokatas?
pl – Słuchaj-no, ***Mikołaj***! – ciszej, niż wprzódy mówić zaczął – a wielmi dorohij hetyj hadwokat? (a czy bardzo drogi ten adwokat?).

[19] lt – Ponaiti! ***Steponėl***!… Aš nieko prieš tave neturiu…
pl – Paniczu! *Stefanku*! ja na ciebie nie hniewna (nie gniewam się)…

[20] lt – Eisiu jau, ***Motin*** ir *ponia brangi* – tarė – eisiu jau; aš paukštis keleivis…
pl – *Matko* a *pani*! – rzekł – pójdę już; jam ptak wędrowny…

W większości przykładów [13–20] stwierdza się równoległe użycie wołacza w obu językach. Tylko w przykładzie [18] polskiej formie podstawowej *Mikołaj* odpowiada litewska w wołaczu *Mikalojau*. Pewnej uwagi wymagają przykłady [13–15]. Polskiej formie *Jasiuk* w litewskim przekładzie odpowiada raz forma niedeminutywna *Jonai*, raz deminutywna – *Joneli*. Nie to jednak powinno przykuć uwagę czytelnika, lecz sama postać polskiej formy *Jasiuk*, która jest całkowicie zgodna z litewskim wołaczem, por. litew. nom. sg. *Jasiukas* i voc. sg. *Jasiuk*. Warto przy okazji zwrócić uwagę na tę postać litewskiego wołacza. Jest to czysty temat. Tu należy wyjaśnić, że w normalizowanej litewszczyźnie preferowana forma wołacza posiada fleksję, por. litew. nom. sg. *Jasiuk-as* i voc. sg. *Jasiuk-**ai*** i in. W rejestrze mówionym (co jest odzwierciedlone w beletrystyce i w nagraniach gwarowych) obserwuje się użycie samego tematu jako wołacza, por. już wyżej przytaczane formy litewskie *vaikuti* 'dziecko (voc.)' [5], *berneli* 'chłopcze' [5], *broli* 'bracie' [6, 10], *vyreli*[30] '~ chłopie' [11], *Steponėl* 'Stefanku' [19], *motin* 'matko' [20] i in.

Przykłady [6–20] odzwierciedlają stan języków polskiego i litewskiego końca XIX wieku. Ponadto, należy zauważyć, twórcy tych utworów byli w mniejszym lub większym stopniu związani z areałem języka litewskiego oraz poruszali tematykę „litewską". Aby wykluczyć wpływ języka litewskiego na pisarza, warto więc sięgnąć po inny przykład utworu tamtego okresu. Niech będzie to również dostępne w korpusie IS PAN – Clarin-PL dzieło Adama Asnyka *Kiejstut* (1843), poety i dramatopisarza, który, w odróżnieniu od wcześniej wymienionych pisarzy

[30] Wołacz utworzony od formy deminutywnej *vyr-el-is* < *vyr-as* 'mężczyzna'. Wypada podkreślić, że utworzenie formy deminutywnej w języku litewskim jest zabiegiem prostym. Nie da się tego powiedzieć o języku polskim, czego przykładem może być trudność urobienia zdrobnienia od leksemu *mężczyzna*.

polskich, nie był związany z areałem litewskim, Okazuje się, że również Adam Asnyk konsekwentnie stosuje formy wołacza, por. pol. *wielki kniaziu* – litew. *didi*[31] *kunigaikšti*, pol. *panie* – litew. *viešpatie*, pol. *bracie* – litew. *broli*, pol. *psie nikczemny* – litew. *šunie*, pol. *ojcze złoty* – litew. *tėveli*, pol. *mój staruszku* – litew. *mano seni*, pol. *ojcze* – litew. *tėve Kęstuti*, pol. *Niemcze* – litew. *vokieti*, pol. *synu* – litew. *sūnau*, pol. *Witoldzie* – litew. *Vytautai*, pol. *starcze* – litew. *seni*, pol. *książę*[32] *Kiejstucie* – litew. *kunigaikšti Kęstuti*, pol. *dziecko*[33] – litew. *vaikeli* i in. Polskiego badacza mogą zainteresować litewskie formy przymiotnika w wołaczu (stary temat), por. pol. *okrutny* – litew. *beširdi* (adj. nom. sg. *beširdis*), pol. *bezczelny* – litew. *begėdi* (adj. nom. sg. *begėdis*) i in.

Zestawienie przykładów polskich i litewskich zaczerpniętych z dzieł datowanych na wiek XIX z tymi pochodzącymi z drugiej połowy XX wieku uświadamia użytkownikom polszczyzny fakt, że liczba użyć wołacza w języku polskim na przestrzeni niespełna jednego wieku zdecydowanie zmalała. Można też odnieść wrażenie, że użycie polskiego wołacza w rejestrze mówionym, w listach i pismach oficjalnych jest oznaką szacunku do odbiorcy. Tym samym pierwotna funkcja wołacza została przewartościowana.

Rusycystom znane jest zjawisko użycia tematu leksemu w funkcji wołacza, por. ros. *мам*, *пап*, *Ир*, *Лен*, *Марин*, *ребят* (pl.). W zestawieniu tych form z litewskimi, zauważa się ten sam mechanizm budowy wołacza. W języku rosyjskim te formy określa się terminem nowego wołacza (ros. *новозвательный падеж*) lub też współczesnego wołacza (ros. *современный звательный падеж*) (por. Кронгауз, 1999; Полонский, 2002; Супрун, 2001 i in.). W języku rosyjskim, w odróżnieniu od litewskiego, użycie samego tematu leksemu w funkcji wołacza jest jedynym wykładnikiem tej kategorii. Pierwotna forma wołacza w języku rosyjskim stała się nieproduktywna, zanikła. Fakt użycia tematu leksemu w rosyjskim i litewskim jako wołacza, równoległe funkcjonowanie dwóch form wołacza w litewskim (sam temat lub fleksja) oraz zachowanie form wołacza w bułgarskim mogą potwierdzać tezę Kuryłowicza o konieczności wykreślenia wołacza z wykazu przypadków w językach indoeuropejskich[34].

---

[31] Przymiotnik *didis* 'wielki' w formie wołacza.

[32] Forma wołacza *Kiejstucie* warunkuje użycie formy rodzaju nijakiego *książę* również w wołaczu.

[33] W tym przypadku (wyrażenie proste fundowane przez rzeczownik rodzaju nijakiego *dziecko*) możemy przyjąć, zważywszy na konsekwentne stosowanie formy wołacza przez A. Asnyka w tym utworze, że jest to również forma wołacza.

[34] Rozważania tu zasygnalizowane można podciągnąć pod badania zwane w anglojęzycznej literaturze terminami *corpus-based approach* i *corpus-supported approach*.

Trudno to sobie wyobrazić, a jednak utrata formy wołacza w języku rosyjskim wpłynęła na postrzeganie przez Bułgarów tejże w swoim języku. Takie są fakty. Do odzyskania niepodległości przez Bułgarię na początku XX wieku przyczyniła się carska Rosja. Następnie w latach 20. XX wieku do Bułgarii przybywali dobrze wykształceni emigranci rosyjscy, których obdarzono dużym szacunkiem. Fakt, nie jest to naukowy wywód, lecz jedynie powtórzenie wielokrotnie słyszanego w Bułgarii wyjaśnienia[35], jakoby w podzięce Rosji naród bułgarski widział w tym kraju, Rosjanach i ich języku wzorzec wszelkich wartości. Skoro w języku rosyjskim wołacz zanikł, to również nie jest on konieczny w bułgarskim, zwłaszcza gdy użycie innych przypadków w języku bułgarskim zanikło. Stosowanie form wołacza w bułgarskim rzekomo miało być oznaką człowieka niewykształconego, o niższej kulturze osobistej. O różnym rozwoju wołacza w rosyjskim i bułgarskim pisał między innymi L. Andrejchin (Андрейчин, 1978, s. 244).

Współautor tego artykułu przypomina sobie sytuację z posiedzenia Polsko-Bułgarskiego Zespołu do spraw Gramatyki Konfrontatywnej (przełom dekad 80./90. XX wieku), gdy bezpośredni zwrot prof. Violetty Koseskiej do bułgarskiego kolegi prof. Jordana Penčewa *Йордане* wywołał zagorzałą dyskusję na temat wołacza w języku bułgarskim. Postronnemu obserwatorowi spór, który wywiązał się między Bułgarami-bułgarystami, wydawał się wyłącznie emocjonalny a nie merytoryczny. Nie dyskutowano, czy wołacz należy uznać za jeden z przypadków, czy też za oddzielną kategorię, jak to uważali L. Andrejchin (Андрейчин, 1952) czy P. Pashov (Пашов, 1989). Nie mówiono, jakie są ograniczenia w tworzeniu form wołacza, kiedy dochodzi do neutralizacji formy podstawowej i wołacza, jaki jest wpływ form wołacza na *-o* (typu bułg. *Марийо*), które uważa się za deprecjonujące, na rugowanie wołacza z systemu i in. Czyżby o wciąż niezamkniętej kwestii wołacza w języku bułgarskim świadczyć miało błędne tagowanie form wołacza w BgTaggerze (*BgTagger*, b.d.)[36]? A może twórczy BgTaggera założyli, że wołacz jest na tyle marginalnym zjawiskiem w języka bułgarskim, że nie zachodzi potrzeba uczenia narzędzia rozpoznawania tych form? Tylko niektóre formy BgTagger interpretuje właściwie, np. *Господине* czy *Петре*, por.

[21]				
господин	N	господин	NCMsom	(słowoforma zgodna z formą bazową)
господине	NH	Господин	NHMsvm	(zidentyfikowana forma wołacza)
[22]				
Петър	NH	Петър	NHMsom	(słowoforma zgodna z formą bazową)
Петре	NH	Петър	NHMsvm	(zidentyfikowana forma wołacza)

[35] Takie sądy słyszeliśmy z ust bułgarskich historyków i językoznawców.

[36] *Тагерът „BgTagger" за български език* (BgTagger (b.d.).

Inne zaś zostają błędnie opisane lub nierozpoznane, por.

[23]

човек	N	човек	NCMsom (słowoforma zgodna z formą bazową)
човече	N	човече	NCNson (słowoforma zgodna z formą bazową)
ветре	N	ветре	N
брате	N	брате	N
другарю	N	другарю	N
приятелю	N	приятелю	N
бабо	N	бабо	N
майко	N	майко	N
жѐно	A	жѐно	A

W dialogach filmowych, zarówno polskich, jak i bułgarskich, użycie form wołacza jest sporadyczne, por.

[24] pl – Chcesz przeznaczenia, *dupku*?
bg – Искаш съдба, *задник*?

[25] pl – *Christine*!
bg – *Кристин*.

[26] pl. – Przesuń się, *Gus*.
bg – Мърдай, *Гъз*.

[27] pl – Hej, *Murphy*.
bg – Хей, *Мърфи*.

[28] pl – *Jack*, jak leci?
bg – *Джак*, как е хавата?

[29] pl – To jest argument, *Hanson*.
bg – Т'ва е предателство, *Хенсън*.

Możliwe, że obcego pochodzenia nazwy utrudniają utworzenie formy wołacza w obu językach. Do analogicznego wniosku doszła między innymi I. Mankova (Манкова, 2016, s. 12), zestawiając użycie wołacza w bułgarskim i czeskim. Zauważmy, że w języku litewskim ten problem nie występuje, por. wyżej podane przykłady [1–4]. Można też wysnuć przypuszczenie, że autorami tłumaczeń list dialogowych mogą być osoby, które nie wykazują należytej dbałości o język.

We współczesnych utworach beletrystycznych formy wołacza pojawiają się z większą niż w dialogach filmowych regularnością, por.[37]

[37] Анжел Вагенщайн, *Далеч от Толедо* (*Аврам Къркача*) (2011). Angel Wagenstein, *Daleko od Toledo czyli rzecz o Abrahamie Pijanicy* (2011, tłumaczka Kamelia Mincheva-Gospodarek).

[30] bg Какъв манастир, *господи* – толкова човешки и по селски интимен […]
pl Jaki to monastyr? *Boże*, taki ludzki i wieśniaczo intymny […]

[31] bg *Боже господи*, той е на хиляда години, този стар циганин, а с унизено ласкателство ме нарича „*бащице*"!
pl *Mój Boże*, on ma tysiąc lat, ten stary Cygan, a z upokarzającym pochlebstwem nazywa mnie „*ojczulku*".

[32] bg – Бъди благословен, *Мануше*.
pl – Bądź błogosławion, *Manusza*.

[33] bg – *Боже господи*, Аракси… Аракси Вартанян!
pl – *Mój Boże*, Araksi… Araksi Wartanian!

6. Podsumowanie

Europejska infrastruktura Clarin-ERIC systematycznie rozwija się. Rozproszone zasoby (wcześniej powstałe i nowo powstające) zostają połączone w jedną spójną całość. Polskie konsorcjum Clarin-PL przede wszystkim rozwija zasoby i narzędzia dla języka polskiego. Zespół IS PAN opracowuje wielojęzyczne zasoby języków słowiańskich i bałtyckich. Do tej pory opublikowano następujące korpusy: *Polish-Bulgarian Parallel Corpus* (D. Roszko i in., 2018b), *Polish-Bulgarian-Russian Parallel Corpus* (Kisiel i in., 2016), *Polish-Lithuanian Parallel Corpus* (Roszko & Roszko, 2016b), *Polish-Lithuanian Parallel Corpus „2"* (Roszko & Roszko, 2018b), *Polish-Russian Parallel Corpus* (R. Roszko i in., 2018a), *Polish-Ukrainian Parallel Corpus* (Roszko R. i in., 2018b). W roku 2021 udostępnione zostaną korpusy: litewsko-bułgarski, litewsko-rosyjski, litewsko-ukraiński, bułgarsko-rosyjski, bułgarsko-ukraiński oraz rosyjsko-ukraiński. Na rok 2024 przewidziana jest publikacja quasi-referencyjnych korpusów równoległych ręcznie zrównoleglonych i oznakowanych: polsko-bułgarskiego, polsko-litewskiego, polsko-rosyjskiego i polsko-słoweńskiego.

Zastosowania uporządkowanych zasobów językowych, a takimi bez wątpienia są opisane wielojęzyczne korpusy równoległe, są szerokie. Odbiorcami i użytkownikami korpusów są nie tylko badacze szeroko rozumianych nauk humanistycznych i społecznych, lecz również wykładowcy, lektorzy, nauczyciele, osoby uczące się języków. Najnowsze zastosowania korpusów są związane z budową sztucznej inteligencji. Przeprowadzone wstępne analizy użyć wołacza w językach bułgarskim, polskim i litewskim mogą być przykładem na różne sposoby zastosowania korpusów w badaniach. Jednym z tych zastosowań jest ilustracja opisywanego zjawiska przykładami z korpusu (są to tzw. ang. *corpus-illustrated*

/ *corpus-informed approach*). Jest to podstawowy sposób zastosowania korpusów w pracy badawczej. Drugim, zyskującym na popularności, zastosowaniem korpusów są badania prowadzone na materiale korpusowym. Badacz *a priori* formułuje hipotezy, które następnie na materiale korpusowym potwierdza lub neguje (są to tzw. ang. *corpus-based* / *corpus-supported approach*). Trzecim z zastosowań jest prowadzenie na zasobach korpusowych badań od podstaw, których celem jest budowa teorii (są to tzw. ang. *corpus-driven approach*).

Analiza użyć wołacza w językach bułgarskim, litewskim, polskim (z odniesieniami do rosyjskiego) skłania do wniosków, że samego zjawiska wołacza nie należy ściśle łączyć z morfologiczną kategorią przypadka. Wczesny zanik form „starego" wołacza w rosyjskim nie doprowadził do zaniku systemu deklinacji w tymże języku, a „nowy" rosyjski wołacz jest pozbawiony fleksji. W języku bułgarskim możemy mówić o odwrotności zachodzących w języku rosyjskim procesów. System deklinacji rzeczownika zanikł, jednak wołacz pozostał. To, że od blisko stu lat pewne grupy usilnie dążą do wyrugowania wołacza z systemu bułgarszczyzny, tylko potwierdza fakt, że wołacz funkcjonuje niezależnie od systemu deklinacyjnego w danym języku. W przypadku języka litewskiego widzimy, że wołacz występuje tam w dwóch wersjach: bez fleksji i z fleksją. W wyniku prac nad normalizacją języka litewskiego formy fleksyjne uzyskały status form dominujących. Formy bez fleksji, które częściej były zaświadczane w utworach beletrystycznych początku XX wieku, dzisiaj mają status dopuszczalnej formy wołacza.

Badania kontrastywne prowadzone w oparciu o wielojęzyczne korpusy równoległe nie pozwalają skutecznie rozstrzygnąć problemu funkcjonowania wołacza. Jak wiadomo, wołacz jest charakterystyczny dla rejestru mówionego. Zasoby włączone do korpusów równoległych (np. do polsko-bułgarskiego i in.) nie zawierają takich danych, bowiem nie są one dostępne. W korpusach Clarin-PL namiastkę rejestru mówionego stanowią dialogi filmowe.

BIBLIOGRAFIA

BgTagger: Тагерът „BgTagger" за български език. (b.d.). Department of Computational Linguistic IBL-BAS http://dcl.bas.bg/dclservices/index.php

BTB-Pipe: BulTreeBank. (b.d.). http://bultreebank.org/en/clark/

Clarin-PL. (b.d.). *Polska infrastruktura Clarin*. http://clarin-pl.eu/

Cvrček, V., & Richterová, O. (Red.). (2020). *Příručka ČNK: Český národní korpus (January 1, 1970, 00:00 GMT)*. Pobrano 17 czerwca 2020, z http://wiki.korpus.cz/doku.php?id=citation&rev=0

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2009a). 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* (ss. 1–8). Borovets.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2009b). Bulgarian-Polish-Lithuanian Corpus: Problems of development and annotation. W T. Erjavec (Red.), *Research infrastructure for digital lexicography: Mondilex Fifth Open Workshop: Ljubljana, Slovenia, October 14–15, 2009, Proceedings of the 12th International Multiconference Information Society 2009* (ss. 72–86). Department of Knowledge Technologies; Jožef Stefan Institute.

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*, *2010*(10), 217–239. https://doi.org/10.11649/cs.2010.013

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2014). Trilingual aligned corpus: Current state and new applications. *Cognitive Studies | Études cognitives*, *2014*(14), 13–20. https://doi.org/10.11649/cs.2014.002

Dimitrova, L., Pavlov, R., Simov, K., & Sinapova, L. (2005). Bulgarian MULTEXT-East Corpus: Structure and content. *Cybernetics and Information Technologies*, *5*(1), 67–73.

Duszkin, M. (2010). *Wykładniki przybliżoności adnumeratywnej w języku polskim i rosyjskim*. Instytut Slawistyki Polskiej Akademii Nauk.

Karolak, S. (2008). *Gramatyka konfrontatywna bułgarsko-polska: T. 8. Semantyczna kategoria aspektu*. Instytut Slawistyki Polskiej Akademii Nauk.

Kisiel, A., Koseska-Toszewa, V., Kotsyba, N., Satoła-Staśkowiak, J., & Sosnowski, W. (2016). *Polish-Bulgarian-Russian Parallel Corpus: CLARIN-PL digital repository*. https://clarin-pl.eu/dspace/handle/11321/308

Koeva, S., & Genov, A. (2011). Bulgarian Language Processing Chain. W *Proceedings of the Workshop on the Integration of Multilingual Resources and Tools in Web Applications*, 26 September 2011. Association for Computational Linguistics.

KonText. (b.d.). *KonText – Corpus Query Interface*. https://kontext.clarin-pl.eu/run.cgi/first_form

Korytkowska, M. (1992). *Gramatyka konfrontatywna bułgarsko-polska: T. 5. Typy pozycji predykatowo-argumentowych*. Instytut Slawistyki Polskiej Akademii Nauk.

Korytkowska, M. (2004). *Gramatyka konfrontatywna bułgarsko-polska: T. 6, cz. 4. Modalność interrogatywna*. Instytut Slawistyki Polskiej Akademii Nauk.

Korytkowska, M., & Roszko, R. (1997). *Gramatyka konfrontatywna bułgarsko-polska: T. 6, cz. 2. Modalność imperceptywna*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Toszewa, V. (2006). *Gramatyka konfrontatywna bułgarsko-polska: T. 7. Semantyczna kategoria czasu*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2009). *Polsko-bułgarska gramatyka konfrontatywna*. Wydawnictwo Akademickie Dialog.

Koseska-Toszewa, V., Maldžieva, V., & Penčev, J. (1995). *Gramatyka konfrontatywna bułgarsko-polska: T. 6, cz. 1. Modalność: Teoretyczne problemy opisu*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Toszewa, V. & Mazurkiewicz, A. (2010). *Time flow and tenses*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Toszewa, V., & Roszko, R. (2015). On semantic annotation in CLARIN-PL parallel corpora. *Cognitive Studies | Études cognitives*, *15*, 211–236. DOI: https://doi.org/10.116409/cs.2015.016

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

Kuryłowicz, J. (1949). Le problème du classement des cas. *Biuletyn Polskiego Towarzystwa Językoznawczego*, *1949*(9), 20–43.

Kuryłowicz, J. (1968). *O rozwoju kategorii gramatycznych*. Wydawnictwo Naukowe PWN.

LaTeX. (b.d.). *LaTeX: A document preparation system*. https://www.latex-project.org/

Linguee. (b.d.). *Tłumacz i wyszukiwarka zasobów dwujęzycznych*. https://www.linguee.com/?chooseDomain=1

Ljubešić, N., Osenova, P., & Simov, K. (2020). *The CLASSLA-StanfordNLP model for named entity recognition of standard Bulgarian 1.0*. http://doi.org/10.18653/v1/W19-3704

Machálek, T. (2020). KonText: Advanced and flexible corpus query interface. W *Proceedings of LREC 2020* (ss. 7005–7010).

Maldjieva, V. (2009). *Gramatyka konfrontatywna bułgarsko-polska: T. 9. Słowotwórstwo*. Instytut Slawistyki Polskiej Akademii Nauk.

Maldžieva, V. (2003). *Gramatyka konfrontatywna bułgarsko-polska: T. 6, cz. 3. Modalność: Hipotetyczność, irrealność, optatywność i imperatywność, warunkowość*. Instytut Slawistyki Polskiej Akademii Nauk.

MorfoLema. (b.d.). *Analizator morfologiczny języka litewskiego*. http://donelaitis.vdu.lt/MorfoLema/Apie.htm

MULTEXT-East. (b.d.). *Multilingual Text Tools and Corpora for Central and Eastern European Languages. MULTEXT-East Home Page*. http://nl.ijs.si/ME/

ParaConc. (b.d.). *Przeglądarka wielojęzycznych zasobów*. http://www.athel.com/para.html

Roszko, D. (2006). *Funkcjonalne odpowiedniki litewskiego perfectum w litewskiej gwarze puńskiej i w języku polskim*. Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Roszko, D. (2015). *Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej: Na tle literackich języków polskiego i litewskiego*. Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

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 V. Koseska-Toszewa, L. Dimitrova, & R. Roszko (Red.), *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* (ss. 145–158). Institute of Slavic Studies, Polish Academy of Sciences.

Roszko, D., & Roszko, R. (2016a). 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* (ss. 119–132). Instytut Lingwistyki Stosowanej. http://repozytorium.ceon.pl/bitstream/handle/123456789/9717/07_Roszko_Roszko.pdf?sequence=1&isAllowed=y

Roszko, D., & Roszko, R. (2016b). *Polish-Lithuanian Parallel Corpus: CLARIN-PL digital repository*. http://hdl.handle.net/11321/309

Roszko, D., & Roszko, R. (2018a). Polsko-litewskie korpusy IS PAN i CLARIN-PL. W N. Birgiel & D. Roszko (Red.), *Prace bałtystyczne: T. 7. Język – Literatura – Kultura* (ss. 185–205). Uniwersytet Warszawski.

Roszko, D., & Roszko, R. (2018b). *Polish-Lithuanian Parallel Corpus „2": CLARIN-PL digital repository*. http://hdl.handle.net/11321/539

Roszko, D., Roszko, R., & Sosnowski, W. (2018a). Polsko-bułgarskie korpusy IS PAN i CLARIN-PL. *Slavica Lodziensia*, *2*, 59–70.

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

Roszko, R. (2004). Semantyczna kategoria określoności/nieokreśloności w języku litewskim (w zestawieniu z językiem polskim). Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Roszko, R. (2009). Morphosyntactic specifications for Polish. Theoretical foundations. Description of morphosyntactic markers for Polish nouns within MULTEXT-East morphosyntactic specifications (Version 3.0 May 10th, 2004). W L. Dimitrova & R. Garabik (Red.), *Metalanguage and encoding scheme design for digital lexicography. MONDILEX Third Open Workshop. Bratislava, Slovakia, 15–16 April, 2009. Proceedings* (ss. 140–149). Ľ. Štúr Institute of Linguistics.

Roszko, R. (2019). Korpusy wielojęzyczne + przeglądarka korpusowa Kontext. W *Warsztaty CLARIN-PL w praktyce badawczej (UMCS LUBLIN)*. https://nextcloud.clarin-pl.eu/index.php/s/bL6hAv4RyB811F5#pdfviewer

Roszko, R. (2020). *Korpusy wielojęzyczne: polsko-slawistyczno-bałtystyczne*. https://www.youtube.com/watch?v=LcDuZD57mto

Satoła-Staśkowiak, J. (2010). *Polsko-bułgarskie odpowiedniości przekładowe czasów przeszłych*. Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Satoła-Staśkowiak, J. & Koseska-Toszewa, V. (2014). Współczesny słownik bułgarsko-polski – zeszyt I. Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Wstęp. (1984). Projekt gramatyki konfrontatywnej bułgarsko-polskiej i serbskochorwacko-polskiej: Wstęp. W K. Polański (Red.), *Studia polsko-południowosłowiańskie*. Zakład Narodowy im. Ossolińskich.

Андрейчин, Л. (1952). Към въпроса за аналитичния характер на съвременния български език. *Български език*, *1952*(1–2), 20–35.

Андрейчин, Л. (1978). Към въпроса за аналитичния характер на съвременния български език. W П. Пашов (Red.), *Помагало по българска морфология: Имена* (ss. 238–254). Наука и изкуство.

Бояджиев, Т., Стоянов, С., & Попов, К. (Red.). (1983). *Граматика на съвременния български книжовен език: Т. 2. Морфология*. Издателство на Българска академия на науките.

Гугуланова, И., Шимански, М., & Баракова, П. (1993). *Българско-полска съпоставителна граматика: Т. 4. Семантичната категория комуникант*. Издателство на Българска академия на науките.

Косеска-Тошева, В., & Гаргов, Г. (1990). *Българско-полска съпоставителна граматика: Т. 2. Семантичната категория определеност/неопределеност*. Издателство на Българска академия на науките.

Кронгауз, М. А. (1999). Обращения как способ моделирования коммуникативного пространства. W *Логический анализ языка: Образ человека в культуре и языке* (ss. 130–131). Москва.

Крумова-Цветкова, Л., & Рошко, Р. (1994). *Българско-полска съпоставителна граматика: Т. 3, cz. 1. Семантичната категория количество*. Издателство на Българска академия на науките.

Манкова, И. (2016). Употреба на звателни форми като обръщение в чешкия и българския език. *Съпоставителни изследвания*, *41*(*3*), 5–21,

Пашов, П. (1989). *Практическа българска граматика*. Народна просвета.

Петрова-Вашилевич, А., & Чоролеева, М. (1994). *Българско-полска съпоставителна граматика: Т. 3, cz. 2. Семантичната категория степен*. Издателство на Българска академия на науките.

Полонский, А. В. (2002). Эготив, вокатив, номинатив: Субъект и падежная парадигма. *Русский язык за рубежом*, *2002*(3), 27–35. http://gramota.ru/biblio/magazines/ryzr/rzr2001-03/28_197

Супрун, В. И. (2001). Антропонимы в вокативном употреблении. Известия Уральского государственного университета, *20*, 92–96. http://www.philology.ru/linguistics2/suprun_v-01.htm

BIBLIOGRAPHY (TRANSLITERATION)

Andreĭchin, L. (1952). Kŭm vŭprosa za analitichniia kharakter na sŭvremenniia bŭlgarski ezik. *Bŭlgarski ezik*, *1952*(1–2), 20–35.

Andreĭchin, L. (1978). Kŭm vŭprosa za analitichniia kharakter na sŭvremenniia bŭlgarski ezik. In P. Pashov (Ed.), *Pomagalo po bŭlgarska morfologiia: Imena* (pp. 238–254). Nauka i izkustvo.

BgTagger: Тагерът „BgTagger" за български език. (n.d.). Department of Computational Linguistic IBL-BAS http://dcl.bas.bg/dclservices/index.php

Boiadzhiev, T., Stoianov, C., & Popov, K. (Eds.). (1983). *Gramatika na sŭvremenniia bŭlgarski knizhoven ezik: Vol. 2. Morfologiia*. Izdatelstvo na Bŭlgarska akademiia na naukite.

BTB-Pipe: BulTreeBank. (b.d.). http://bultreebank.org/en/clark/

Clarin-PL. (n.d.). *Polska infrastruktura Clarin*. http://clarin-pl.eu/

Cvrček, V., & Richterová, O. (Ed.) (2020). *Příručka ČNK: Český národní korpus (January 1, 1970, 00:00 GMT)*. Retrieved June 17, 2020, from http://wiki.korpus.cz/doku.php?id=citation&rev=0

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2009a). Bulgarian-Polish-Lithuanian Corpus: Current development. In C. Vertan, S. Piperidis, E. Paskaleva, & M. Slavcheva (Eds.), *International workshop: Multilingual resources, technologies and evaluation for Central and Eastern European languages, held in conjunction with the International Conference RANLP-2009: Proceedings* (pp. 1–8). Borovets.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2009b). Bulgarian-Polish-Lithuanian Corpus: Problems of development and annotation. In T. Erjavec (Ed.), *Research infrastructure for digital lexicography: Mondilex Fifth Open Workshop: Ljubljana, Slovenia, October 14–15, 2009, Proceedings of the 12th International Multiconference Information Society 2009* (pp. 72–86). Department of Knowledge Technologies; Jožef Stefan Institute.

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*, *2010*(10), 217–239. https://doi.org/10.11649/cs.2010.013

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2014). Trilingual aligned corpus: Current state and new applications. *Cognitive Studies | Études cognitives*, *2014*(14), 13–20. https://doi.org/

Dimitrova, L., Pavlov, R., Simov, K., & Sinapova, L. (2005). Bulgarian MULTEXT-East Corpus: Structure and content. *Cybernetics and Information Technologies*, *5*(1), 67–73.

Duszkin, M. (2010). *Wykładniki przybliżoności adnumeratywnej w języku polskim i rosyjskim*. Instytut Slawistyki Polskiej Akademii Nauk.

Gugulanova, I., Shymanski, M., & Barakova, P. (1993). *Bŭlgarsko-polska sŭpostavitelna gramatika: Vol. 4. Semantichnata kategoriia komunikant*. Izdatelstvo na Bŭlgarska akademiia na naukite.

Karolak, S. (2008). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 8. Semantyczna kategoria aspektu*. Instytut Slawistyki Polskiej Akademii Nauk.

Kisiel, A., Koseska-Toszewa, V., Kotsyba, N., Satoła-Staśkowiak, J., & Sosnowski, W. (2016). *Polish-Bulgarian-Russian Parallel Corpus: CLARIN-PL digital repository*. https://clarin-pl.eu/dspace/handle/11321/308

Koeva, S., & Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceedings of the Workshop on the Integration of Multilingual Resources and Tools in Web Applications*, 26 September 2011. Association for Computational Linguistics.

KonText. (n.d.). *KonText – Corpus Query Interface*. https://kontext.clarin-pl.eu/run.cgi/first_form

Korytkowska, M. (1992). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 5. Typy pozycji predykatowo-argumentowych*. Instytut Slawistyki Polskiej Akademii Nauk.

Korytkowska, M. (2004). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 6, pt. 4. Modalność interrogatywna*. Instytut Slawistyki Polskiej Akademii Nauk.

Korytkowska, M., & Roszko, R. (1997). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 6, pt. 2. Modalność imperceptywna*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Tosheva, B., & Gargov, G. (1990). *Bŭlgarsko-polska sŭpostavitelna gramatika: Vol. 2. Semantichnata kategoriia opredelenost/neopredelenost*. Izdatelstvo na Bŭlgarska akademiia na naukite.

Koseska-Toszewa, V. (2006). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 7. Semantyczna kategoria czasu*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2009). *Polsko-bułgarska gramatyka konfrontatywna*. Wydawnictwo Akademickie Dialog.

Koseska-Toszewa, V., Maldžieva, V., & Penčev, J. (1995). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 6, pt. 1. Modalność: Teoretyczne problemy opisu*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Toszewa, V. & Mazurkiewicz, A. (2010). *Time flow and tenses*. Instytut Slawistyki Polskiej Akademii Nauk.

Koseska-Toszewa, V., & Roszko, R. (2015). On semantic annotation in CLARIN-PL parallel corpora. *Cognitive Studies | Études cognitives*, *15*, 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*, *51*, 191–217. DOI: https://doi.org/10.11649/sfps.2016.011

Krongauz, M. A. (1999). Obrashcheniia kak sposob modelirovaniia kommunikativnogo prostranstva. W *Logicheskiĭ analiz iazyka: Obraz cheloveka v kul'ture i iazyke* (pp. 130–131). Moskva.

Krumova-TSvetkova, L., & Roshko, R. (1994). *Bŭlgarsko-polska sŭpostavitelna gramatika: Vol. 3, pt. 1. Semantichnata kategoriia kolichestvo*. Izdatelstvo na Bŭlgarska akademiia na naukite.

Kuryłowicz, J. (1949). Le problème du classement des cas. *Biuletyn Polskiego Towarzystwa Językoznawczego*, *1949*(9), 20–43.

Kuryłowicz, J. (1968). *O rozwoju kategorii gramatycznych*. Wydawnictwo Naukowe PWN.

LaTeX. (n.d.). *LaTeX: A document preparation system*. https://www.latex-project.org/

Linguee. (n.d.). *Tłumacz i wyszukiwarka zasobów dwujęzycznych*. https://www.linguee.com/?chooseDomain=1

Ljubešić, N., Osenova, P., & Simov, K. (2020). *The CLASSLA-StanfordNLP model for named entity recognition of standard Bulgarian 1.0*. https://doi.org/

Machálek, T. (2020). KonText: Advanced and flexible corpus query interface. In *Proceedings of LREC 2020*, (pp. 7005–7010).

Maldjieva, V. (2009). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 9. Słowotwórstwo*. Instytut Slawistyki Polskiej Akademii Nauk.

Maldžieva, V. (2003). *Gramatyka konfrontatywna bułgarsko-polska: Vol. 6, pt. 3. Modalność: Hipotetyczność, irrealność, optatywność i imperatywność, warunkowość*. Instytut Slawistyki Polskiej Akademii Nauk.

Mankova, I. (2016). Upotreba na zvatelni formi kato obrŭshtenie v cheshkiia i bŭlgarskiia ezik. *Sŭpostavitelni izsledvaniia*, *41*(*3*), 5–21,

MorfoLema. (n.d.). *Analizator morfologiczny języka litewskiego*. http://donelaitis.vdu.lt/MorfoLema/Apie.htm

MULTEXT-East. (n.d.). *Multilingual Text Tools and Corpora for Central and Eastern European Languages. MULTEXT-East Home Page*. http://nl.ijs.si/ME/

ParaConc. (n.d.). *Przeglądarka wielojęzycznych zasobów*. http://www.athel.com/para.html

Pashov, P. (1989). *Prakticheska bŭlgarska gramatika*. Narodna prosveta.

Petrova-Vashilevich, A., & Choroleeva, M. (1994). *Bŭlgarsko-polska sŭpostavitelna gramatika: Vol. 3, pt. 2. Semantichnata kategoriia stepen*. Izdatelstvo na Bŭlgarska akademiia na naukite.

Polonskiĭ, A. V. (2002). Ėgotiv, vokativ, nominativ: Sub″ekt i padezhnaia paradigma. *Russkiĭ iazyk za rubezhom*, *2002*(3), 27–35. http://gramota.ru/biblio/magazines/ryzr/rzr2001-03/28_197

Roszko, D. (2006). *Funkcjonalne odpowiedniki litewskiego perfectum w litewskiej gwarze puńskiej i w języku polskim*. Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Roszko, D. (2015). *Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej: Na tle literackich języków polskiego i litewskiego*. Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

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). In V. Koseska-Toszewa, L. Dimitrova, & R. Roszko (Eds.), *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* (pp. 145–158). Institute of Slavic Studies, Polish Academy of Sciences.

Roszko, D., & Roszko, R. (2016a). Polsko-litewskie korpusy równoległe: Elementy anotacji semantycznej z zakresu modalności możliwościowej i kwantyfikacji zakresowej. In E. Gruszczyńska & A. Leńko-Szymańska (Eds.), *Polskojęzyczne korpusy równoległe / Polish language parallel corpora* (pp. 119–132). Instytut Lingwistyki Stosowanej. http://repozytorium.ceon.pl/bitstream/handle/123456789/9717/07_Roszko_Roszko.pdf?sequence=1&isAllowed=y

Roszko, D., & Roszko, R. (2016b). *Polish-Lithuanian Parallel Corpus: CLARIN-PL digital repository*. http://hdl.handle.net/11321/309

Roszko, D., & Roszko, R. (2018a). Polsko-litewskie korpusy IS PAN i CLARIN-PL. In N. Birgiel & D. Roszko (Eds.), *Prace bałtystyczne: Vol. 7. Język – Literatura – Kultura* (pp. 185–205). Uniwersytet Warszawski.

Roszko, D., & Roszko, R. (2018b). *Polish-Lithuanian Parallel Corpus „2": CLARIN-PL digital repository*. http://hdl.handle.net/11321/539

Roszko, D., Roszko, R., & Sosnowski, W. (2018a). Polsko-bułgarskie korpusy IS PAN i CLARIN-PL. *Slavica Lodziensia*, *2*, 59–70.

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

Roszko, R. (2004). Semantyczna kategoria określoności/nieokreśloności w języku litewskim (w zestawieniu z językiem polskim). Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Roszko, R. (2009). Morphosyntactic specifications for Polish. Theoretical foundations. Description of morphosyntactic markers for Polish nouns within MULTEXT-East morphosyntactic specifications (Version 3.0 May 10th, 2004). In L. Dimitrova & R. Garabik (Eds.), *Metalanguage and encoding scheme design for digital lexicography. MONDILEX Third Open Workshop. Bratislava, Slovakia, 15–16 April, 2009. Proceedings* (pp. 140–149). Ľ. Štúr Institute of Linguistics.

Roszko, R. (2019). Korpusy wielojęzyczne + przeglądarka korpusowa Kontext. In *Warsztaty CLARIN-PL w praktyce badawczej (UMCS LUBLIN)*. https://nextcloud.clarin-pl.eu/index.php/s/bL6hAv4RyB811F5#pdfviewer

Roszko, R. (2020). *Korpusy wielojęzyczne: polsko-slawistyczno-bałtystyczne*. https://www.youtube.com/watch?v=LcDuZD57mto

Satoła-Staśkowiak, J. (2010). *Polsko-bułgarskie odpowiedniości przekładowe czasów przeszłych*. Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Satoła-Staśkowiak, J. & Koseska-Toszewa, V. (2014). *Współczesny słownik bułgarsko-polski – zeszyt I.* Instytut Slawistyki Polskiej Akademii Nauk (Slawistyczny Ośrodek Wydawniczy).

Suprun,V. I. (2001). Antroponimy v vokativnom upotreblenii. Izvestiya Ural'skogo gosudarstvennogo universiteta, *20*, 92–96. http://www.philology.ru/linguistics2/suprun_v-01.htm

Wstęp. (1984). Projekt gramatyki konfrontatywnej bułgarsko-polskiej i serbsko-chorwacko-polskiej: Wstęp. In K. Polański (Ed.), *Studia polsko-południowosłowiańskie*. Zakład Narodowy im. Ossolińskich.

Korpusy wielojęzyczne wkładem Instytutu Slawistyki Polskiej Akademii Nauk w rozwój infrastruktury Clarin-PL. Przykłady analizy korpusowej nad wołaczem

Abstrakt

W artykule opisano budowane przez zespół językoznawców Instytutu Slawistyki Polskiej Akademii Nauk (dalej IS PAN) korpusy wielojęzyczne z węzłowym językiem polskim. W pierwszej części artykułu (podpunkty 1–3) przedstawiono przyczyny, które doprowadziły do rozwoju lingwistyki korpusowej w IS PAN oraz opisano pierwsze powstałe w IS PAN korpusy. W drugiej części artykułu (podpunkt 4) przybliżono charakter infrastruktury Clarin oraz konstruowane w IS PAN na rzecz tej infrastruktury wielojęzyczne korpusy języków słowiańskich i bałtyckich. W części trzeciej artykułu (podpunkt 5) na przykładach ilustrujących użycie form wołacza w językach polskim, bułgarskim i litewskim zaprezentowano różne warianty zastosowania korpusów wielojęzycznych w badaniach kontrastywnych. Omawiane zagadnienia ilustrowano przykładami wyekscerpowanymi z tychże korpusów.

Słowa kluczowe: infrastruktura badawcza Clarin-PL; wielojęzyczne korpusy; badania kontrastywne; badania korpusowe; wołacz

The Contribution of the Institute of Slavic Studies, Polish Academy of Sciences to the Development of the Clarin-PL Infrastructure: Examples of Corpus Analysis of the Vocative Case

Abstract

This article describes multilingual corpora with Polish as the hub language which have been constructed by a team of linguists from the Institute of Slavic Studies, Polish Academy of Sciences. The first part (Sections 1–3) outlines the factors which led to the development of corpus linguistics studies at the Institute and presents its first corpora. Section 4 provides an overview of the Clarin infrastructure and the multilingual corpora of Slavic and Baltic languages which were designed under this framework by a team working at the Institute. Section 5 presents potential applications of multilingual corpora in contrastive studies, using examples of the vocative case in Polish, Bulgarian and Lithuanian. The issues in focus are illustrated with examples extracted from the corpora under discussion.

Keywords: Clarin-PL infrastructure; multilingual corpora; contrastive studies; corpus research; vocative case