https://doi.org/10.25312/2391-5137.13/2019_13rt

Roman Tymoshuk https://orcid.org/0000-0003-1391-164X
Instytut Slawistyki Polskiej Akademii Nauk, Warszawa
roman.tymoshuk@gmail.com

Korpusy równoległe a język i społeczeństwo, czyli o znaczeniu, praktycznym zastosowaniu i perspektywach rozwoju lingwistyki korpusowej (na przykładzie korpusów równoległych polsko-ukraińskiego i polsko-rosyjskiego)

Streszczenie

Artykuł porusza kwestie dotyczące roli lingwistyki korpusowej oraz badań interdyscyplinarnych we współczesnym językoznawstwie. Omówione zostały możliwości zastosowania wielojęzycznych zasobów cyfrowych w badaniach lingwistycznych. Przedstawiono przykłady wykorzystania korpusów równoległych w badaniach nad współczesnym słownictwem i frazeologią języków słowiańskich. Rozważania prowadzą do wniosku, że obecnie, gdy coraz bardziej rośnie zapotrzebowanie na zastosowanie mechanizmów języka naturalnego w systemach informacyjno-komputerowych oraz interakcji człowiek–komputer, konieczny jest rozwój zasobów i narzędzi do przetwarzania języka umożliwiających skuteczne pokonywanie barier językowych. Pozwoli to na bliższą współpracę między badaczami reprezentującymi różne nauki.

Słowa kluczowe: technologie językowe, korpus równoległy, nowe słownictwo, frazeologia, ekwiwalencja międzyjęzykowa

Wstęp

W epoce cyfrowej dane korpusowe stały się nieodłączną częścią badań lingwistycznych. W ostatnim czasie rośnie zainteresowanie lingwistyką korpusową wśród badaczy reprezentujących różne dziedziny, nie tylko językoznawstwo. Pojawiają się nowe projekty, w ramach których powstają narzędzia językowe oraz zasoby informacyjne z zasadniczo

nowymi możliwościami analizy i przetwarzania informacji w języku naturalnym[1]. W rezultacie w lingwistyce wyodrębniła się nowa dziedzina – technologie językowe.

Jednym z kierunków lingwistyki korpusowej jest tworzenie korpusów równoległych, które są wykorzystywane do rozwiązywania różnych zadań, takich jak: tworzenie systemów tłumaczenia maszynowego, budowa bazy pamięci tłumaczeniowej, metodyka nauczania języków obcych. Korpusy równoległe oraz ich wyszukiwarki udostępniają lingwistom, nauczycielom języków obcych, tłumaczom, kulturoznawcom i innym badaczom bezcenny i wcześniej niedostępny materiał językowy, który znajduje swoje zastosowanie w wielu dziedzinach, zwłaszcza tam, gdzie dokonuje się zestawienie dwóch lub więcej języków i kultur. Takie narzędzia są niezbędne, ponieważ wiarygodne i potwierdzone empirycznie odpowiedzi na większość pytań językoznawstwa kontrastywnego, w tym leksykologii i leksykografii, mogą być uzyskane wyłącznie na podstawie ogólnodostępnych korpusów równoległych o dużej objętości (por. Dobrovol'skij, 2009; Meyer, 2004).

W ciągu ostatnich lat w Polsce i za granicą powstało i nadal powstaje wiele polskojęzycznych korpusów równoległych. Wśród nich są następujące: polsko-angielski korpus równoległy Paralela, polsko-niemiecki i niemiecko-polski korpus równoległy, polsko-rosyjski i rosyjsko-polski korpus równoległy, polsko-słowacki korpus równoległy, oraz wielojęzyczne: InterCorp, ParaSol, Opus[2]. Celem takich inicjatyw naukowych jest pokonanie barier językowych oraz wspieranie badań humanistycznych w wielokulturowej i wielojęzycznej Europie.

Korpusy równoległe Clarin-PL

W ramach polskiej części infrastruktury naukowej Clarin ERIC[3] zespół Instytutu Slawistyki PAN pracuje nad budową dwujęzycznych korpusów tekstów: polsko-bułgarskiego, polsko-rosyjskiego, polsko-ukraińskiego oraz polsko-litewskiego[4]. Do korpusów równoległych Clarin-PL weszły teksty od końca XX do początku XXI wieku, odzwierciedlające obecny rozwój cywilizacyjny (utwory beletrystyczne, teksty prawne Unii Europejskiej, umowy, dokumentacja techniczna i inne). W ramach planowanych prac przewiduje się rozszerzenie bazy o inne teksty i gatunki. Należy podkreślić, że tworzenie dwu- lub wie-

[1] Według prognoz firmy Seagate i IDC całkowita ilość informacji cyfrowej, wyprodukowanej przez człowieka do 2025 roku wyniesie 175 zettabajtów. To prawie dziesięciokrotny wzrost w porównaniu do 2016 roku. Prawie 30% światowych danych będzie przetwarzane w czasie rzeczywistym, https://www.seagate.com/pl/pl/our-story/data-age-2025/ [dostęp: 15.04.2019].

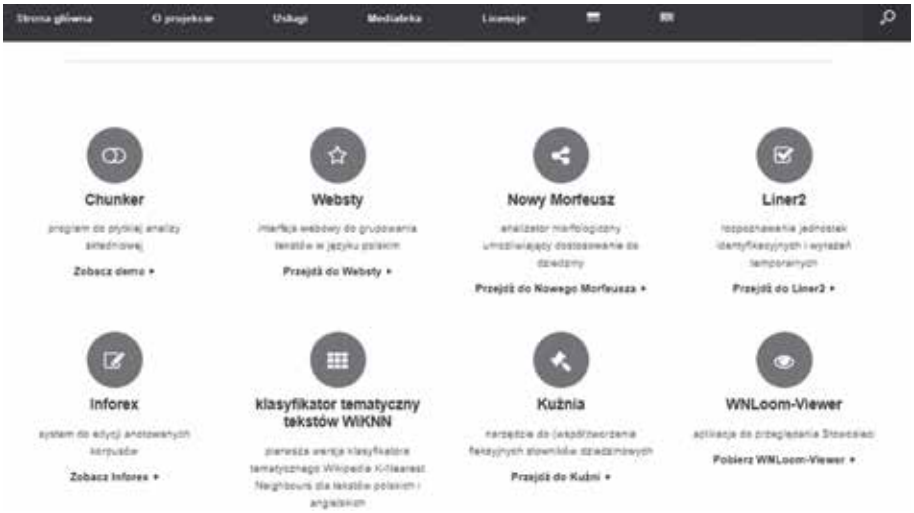
[2] Przegląd zasobów elektronicznych dla poszczególnych języków słowiańskich por. Leńko-Szymańska, Gruszczyńska, 2016; dla języka polskiego, http://clip.ipipan.waw.pl/LRT [dostęp: 15.04.2019].

[3] CLARIN ERIC (Common Language Resources and Technology Infrastructure, European Research Infrastructure Consortium) to ogólnoeuropejska infrastruktura naukowa, część Europejskiej Mapy Drogowej Infrastruktury Naukowej (ESFRI – European Roadmap for Research Infrastructures, European Strategy Forum on Research Infrastructures). Celem CLARIN jest udostępnianie zasobów językowych oraz elektronicznych narzędzi do automatycznego przetwarzania języka naturalnego badaczom we wszystkich dyscyplinach naukowych, a w szczególności z dziedziny nauk humanistycznych i społecznych.

[4] W celu uzyskania pełnego dostępu do korpusów użytkownik powinien zarejestrować się na stronie i skorzystać z narzędzia do wyszukiwania zasobów KonText, https://clarin-pl.eu/dspace/register [dostęp: 15.04.2019].

lojęzycznych zbiorów cyfrowych tego typu jest dość trudnym i bardzo czasochłonnym procesem, ponieważ większość współczesnych tekstów oraz ich tłumaczenia są chronione prawem autorskim i faktycznie każdy tekst musi być zdobywany dwa razy. Oprócz tego na dzień dzisiejszy wiele aspektów prac pod względem technicznym nie podlega automatyzacji i jest wykonywane ręcznie.

Obecnie w ramach sieci Clarin-PL udostępnia się takie zasoby, jak: wyszukiwarka polsko-angielskich anotowanych korpusów równoległych Paralela[5], *Słowosieć* – wordnet języka polskiego[6], *Walenty* – słownik walencyjny języka polskiego[7], *Spokes* – wyszukiwarka danych konwersacyjnych[8], programy do analizy składniowej i morfologicznej, narzędzie do wyznaczania słów kluczowych w tekście, narzędzie do wykrywania obcojęzycznych wtrąceń w tekście i inne[9]. Tego rodzaju zasoby i narzędzia do przetwarzania języka będą przydatne badaczom w analizie dyskursu politycznego, społecznego lub reklamowego.



Rysunek 1. Zasoby i narzędzia Clarin-PL
Źródło: CLARIN-PL, http://clarin-pl.eu/en/home-page/.

Badania nad frazeologią

Korpusy równoległe mogą być wykorzystywane skutecznie w różnych badaniach lingwistycznych, w szczególności w dziedzinie frazeologii porównawczej (por. Sosnowski, Blagoeva, Tymoshuk, 2018; Pędzik, 2016). Korzystając z korpusów, leksykografowie, tłumacze i nauczyciele języka otrzymują bardzo proste i skuteczne narzędzie do gro-

[5] http://paralela.clarin-pl.eu/ [dostęp: 15.04.2019].
[6] http://plwordnet.pwr.wroc.pl/wordnet/ [dostęp: 15.04.2019].
[7] http://walenty.ipipan.waw.pl/ [dostęp: 15.04.2019].
[8] http://spokes.clarin-pl.eu/ [dostęp: 15.04.2019].
[9] http://clarin-pl.eu/pl/uslugi/ [dostęp: 15.04.2019].

madzenia materiału lingwistycznego i sprawdzenia swoich hipotez dotyczących ekwiwalencji międzyjęzykowej. Z punktu widzenia teorii frazeologii, zwłaszcza aspektu konfrontatywnego, pojawia się pytanie o istotę ekwiwalencji międzyjęzykowej. Dmitrij Dobrovol'skij (2015: 26) wyodrębnia następujące aspekty ekwiwalencji: „(a) ekwiwalencja w tłumaczeniu, tzn. relacje między danym idiomem języka *L*1 i jego tłumaczeniem na język *L*2 w pewnym tekście i (b) ekwiwalencja w systemie języka, tzn. relacje między konfrontowanymi idiomami języków *L*1 i *L*2 na poziomie systemowym".

Ekwiwalencja tłumaczeniowa w dużej mierze zależy od kontekstu, dlatego ustalenie pary ekwiwalentów przekładowych jest procesem dość trudnym, wymagającym od badaczy analizy lingwistycznej dużych zasobów tekstów. Takie możliwości dają korpusy równoległe. Oto kilka przykładów wyszukiwania ekwiwalentów przekładowych. Materiał polsko-ukraińskiego korpusu równoległego oraz słowniki lingwistyczne podają informację, że polska jednostka frazeologiczna *wziąć się w garść* stanowi ekwiwalentną parę z ukraińskim frazeologizmem *взяти себе в руки* (por. LAFPU: 137). W korpusie odnotowane są również inne odpowiedniki frazeologiczne w języku ukraińskim: *зібрати себе до купи, взятися за розум, оволодіти собою* itd. (rys. 2). To wskazuje, że w pewnych kontekstach język docelowy wykorzystuje różne środki wyrażania sytuacji. Polsko-rosyjski korpus podaje takie wyniki wyszukiwania polskiego frazeologizmu *wziąć się w garść*: ros. *взять себя в руки, обрести почву под ногами, прийти в себя*, a także *собраться*. Jak widać, często jednostka idiomatyczna języka wyjściowego może mieć mniej idiomatyczne/swobodne odpowiedniki lub odpowiedniki w postaci słowa.

Rysunek 2. Wyniki wyszukiwania dla frazeologizmu *wziąć się w garść* w korpusie PolUkr
Źródło: PolUkr.

Należy podkreślić, że korpus równoległy powinien zawierać wystarczającą liczbę tekstów różnych gatunków, które odzwierciedlają system językowy. Przykładami takich tekstów są literatura piękna, publicystyka, teksty dotyczące polityki i technologii telekomunikacyjnych, teksty naukowo-popularne, teksty zawierające język mówiony itp. Zdaniem twórców korpusów bogatym źródłem słownictwa potocznego są dialogi współczesnych filmów fabularnych i seriali, a także filmy w różnych wersjach językowych (opatrzone napisami lub dubbingowane)[10]. Włączenie takich zasobów do korpusów umożliwia

[10] Po raz pierwszy filmy fabularne jako źródło polskiej i ukraińskiej współczesnej frazeologii zostały wykorzystane podczas tworzenia *Leksykonu aktywnej frazeologii polskiej i ukraińskiej* autorstwa R. Tymoshuka, W. Sosnowskiego, M. Jaskota i Y. Ganoshenki. *Leksykon* zawiera ponad 1000 jednostek frazeologicznych używanych

wyszukiwanie jednostek potocznych, neologizmów, wulgaryzmów i innych jednostek językowych, których zazwyczaj nie zawierają słowniki przekładowe i jednojęzyczne słowniki frazeologiczne.

Do korpusów wielojęzycznych Clarin-PL weszły dialogi/napisy filmowe w języku polskim, ukraińskim i rosyjskim[11]. Takie zasoby stanowią cenny materiał dla językoznawców, ponieważ zawierają jednostki wyrażające ocenę pewnej sytuacji, a także różne stany emocjonalne uczestników komunikacji. Na przykład polsko-ukraiński korpus równoległy podaje takie wyniki wyszukiwania kolokwialnej jednostki frazeologicznej *mam to gdzieś*: ukr. *мені начхати, мені все одно, немає про що хвилюватися*. Dziewięć wyników tego frazeologizmu podaje prawa część polsko-rosyjskiego korpusu: rosyjskie ekwiwalenty *мне все равно, мне плевать, меня не волнует* (rys. 3).

Rysunek 3. Wyniki wyszukiwania dla frazeologizmu *mam to gdzieś* w korpusie PolRU

Źródło: PolRU.

W analizowanych korpusach równoległych znajdujemy wyrażenie *mieć przerąbane*, używane w znaczeniu ‘ktoś znalazł się w sytuacji dla siebie bardzo niekorzystnej, z której w ogóle nie ma wyjścia albo jest je bardzo trudno znaleźć’ (WSJP) (zob. tab. 1). Analiza korpusowa potwierdza również aktywne używanie wulgarnego wariantu tego frazeologizmu *mieć przejebane* (ukr.: *бути в (повній) дупі, бути в лайні*; ros. *быть/оказаться в заднице*)[12]. Warto jednak odnotować, że postrzeganie wyrażeń wulgarnych jest znacznie łagodniejsze, niż można było zaobserwować w przeszłości. Jak zauważa Magdalena Hądzlik-Dudka (2014: 156), coraz częściej daje o sobie znać zjawisko dewulgaryzacji wulgaryzmów, czyli ich obłaskawiania lub nobilitacji.

aktywnie we współczesnych językach polskim i ukraińskim. Zaletą *Leksykonu* jest to, że poza frazeologią ogólną znalazły się w nim też wybrane neologizmy frazeologiczne oraz jednostki, które często nie mają odpowiedników w innym języku, dlatego że odzwierciedlają kulturę danego narodu i jego językowy obraz świata – zob. więcej Sosnowski, Tymoshuk, 2017; https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.1317 [dostęp: 15.04.2019].

[11] Napisy filmowe zawierają również inne korpusy, na przykład korpus równoległy InterCorp v11 w ramach Czeskiego Korpusu Narodowego, https://korpus.cz/ [dostęp: 15.04.2019].

[12] O wzroście częstotliwości stosowania przez użytkowników wulgaryzmów w codziennej komunikacji świadczą wyniki badań *Wulgaryzmy w życiu codziennym* przeprowadzone przez Centrum Badania Opinii Społecznej. Wyniki badań z 2013 roku wskazują, że niemal 80% ankietowanych używało wulgaryzmów. Używanie wulgaryzmów pod wpływem emocji zadeklarowało 65% ankietowanych, https://www.cbos.pl/SPISKOM.POL/2013/K_136_13.PDF [dostęp: 15.04.2019].

Tabela 1. Przykłady użycia jednostki frazeologicznej *mieć przerąbane* w PolUkr i PolRU

PolUkr	„Biedny Byron. Kurewskie szczęście, co? Płakać się chce. Niektórzy to **mają przerąbane**".	„Бідний Байрон. Жахлива річ – удача. Та вже, дивись, не заплач. У деяких все просто жахливо".
PolRU	„Może to miejsce już nie istnieje. O kurczę. **Masz przerąbane**, dupku. Słyszysz? **Przerąbane**. Oto dzień sądu".	„Может, этого места вообще уже нет. Ой, дорогуша. Ты облажался, ублюдок. Слышишь? Облажался. Сегодня день расплаты, сэр".

Podsumowując, należy podkreślić, że przy wystarczającej objętości dane korpusowe umożliwiają określenie frekwencji użycia poszczególnych jednostek frazeologicznych w różnych gatunkach tekstów, ustalenie wariantywności składu leksykalnego frazemu oraz dobór adekwatnych odpowiedników przekładowych dla każdego znaczenia idiomu.

Neologia i korpusy równoległe

Na przełomie XX i XXI wieku społeczność światowa weszła w nową fazę ewolucji – społeczeństwa informacyjnego oraz społeczeństwa wiedzy[13], co oznacza, że informacja zaczęła stanowić podstawę działania gospodarki i odzwierciedlać realia życia społecznego. Dzięki intensywnemu rozwojowi i wdrażaniu nowoczesnych narzędzi komunikacyjnych człowiek z dowolnego miejsca na świecie potrafi natychmiast połączyć się z inną osobą lub źródłem informacji. To ma znaczący wpływ na współczesne języki naturalne, które jako otwarte i dynamiczne systemy ciągle się rozwijają. Obecnie języki słowiańskie intensywnie ewoluują i wśród wszystkich ich poziomów najbardziej dynamicznie rozwija się poziom leksykalno-semantyczny. Dlatego kwestia pojawienia się i funkcjonowania nowych słów pozostaje aktualna zarówno pod względem teoretycznym, jak i praktycznym.

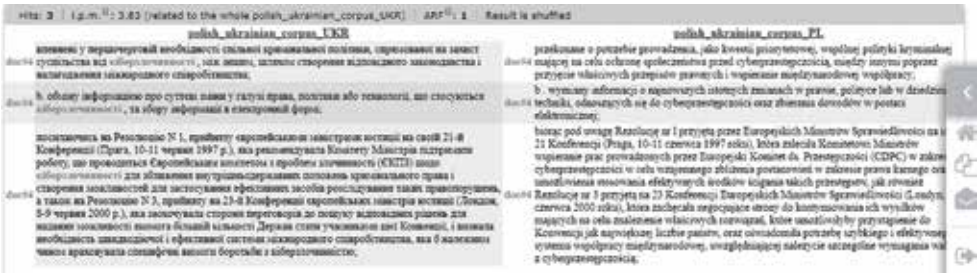
Współczesne języki słowiańskie podlegają uniwersalnym tendencjom do demokratyzacji (bardziej ogólnie rozumianej jako kolokwializacja), intelektualizacji i internacjonalizacji (zob. Popova, 2005; Styshov, 2015). W dobie globalizacji wzrasta liczba kontaktów i interakcja między przedstawicielami różnych grup etnicznych i ich języków, co spowodowało intensyfikację zapożyczeń z języka angielskiego i jego odmiany amerykańskiej. Na temat internacjonalizacji w językach słowiańskich powstało wiele rozpraw (por. np. Waszakowa, 2009, 2011; Blagoeva, 2008; Bozděchová, 2010). Zauważalny jest znaczący wpływ języka potocznego na język literacki, zwłaszcza slangu młodzieżowego i języka subkultur (por. Satoła-Staśkowiak, 2016): pol. *czaić*, *pojechać po bandzie*, *czadowo*, *lajkować*, *fejs*, *apka*, *przehajpować*; ukr. *бути в темі*, *дах поїхав*, *хейтити* (w znacze-

[13] Więcej na temat ewolucyjnych procesów społecznych, a także o informacyjnym podejściu do opisu języka por. Shyrokov, 2017, a także Krztoń, 2015.

niu ‘nienawidzieć’), *зафрендити*, *постити*, *лайфхак*; ros. *крышу сносит*, *хайпануть* (w znaczeniu ‘zdobyć sławę’), *ламповый* (w znaczeniu ‘przyjemny’), *запостить*, *фаннейдж*, *личка*, *левел*, *майнинг-ферма*, *хейтер* (w znaczeniu ‘nieprzyjaciel’)[14]. Neologizmy pochodzenia slangowego coraz częściej służą jako środek ekspresyjnej samorealizacji niż jako znaki przynależności do grupy społecznej. Analizując polszczyznę przełomu XX i XXI wieku, Krystyna Waszakowa (2011: 6–7) stwierdza, że:

> Coraz wyraźniej zaznacza się dominacja komunikacji elektronicznej nad innymi formami komunikowania […]. Zwraca się uwagę na pojawienie się nowego stylu polszczyzny, sytuującego się pomiędzy polszczyzną staranną (pisaną) a potoczną (mówioną) – chodzi o język czatów, esemesów, e-maili, forów internetowych, blogów i in.

Istotnym źródłem nowego słownictwa są wszelkiego rodzaju przemiany społeczne, polityczne i gospodarcze. Takie jednostki najczęściej powstają i funkcjonują w stylu dziennikarsko-publicystycznym, na przykład: ukr. *небесна сотня* (‘patrioci, którzy zginęli podczas akcji protestu na Ukrainie w latach 2013–2014’), *зоряна війна* (‘konflikt międzyplanetarny’), *іти в тінь* (‘prowadzić działalność gospodarczą nielegalnie’), *сіра економіка* (‘nierejestrowana aktywność gospodarcza’); pol. *biała szkoła* (‘zimowa wycieczka szkolna w celach rekreacyjnych i edukacyjnych, trwająca zwykle kilka dni’), *z niższej/wyższej półki* (‘coś niższej/wyższej jakości’), *jechać/jeździć na saksy* (‘wtedy, gdy ktoś wyjeżdża zarabiać za granicą’ – teraz już w nowym znaczeniu, bo nie tylko do Niemiec). Duża grupa neologizmów powstaje w związku z postępem technologicznym społeczeństwa, pojawieniem się nowych koncepcji, wynalazków i innowacji. Materiał lingwistyczny wskazuje, że w trzech językach słowiańskich funkcjonują takie neologizmy terminologiczne jak: pol. *rzeczywistość wirtualna*, *marketing cyfrowy*, *kryptowaluta*, *e-biznes*, *cyberprzestępczość*; ukr. *віртуальна реальність*, *кіберзлочинність*, *цифровий стиль*, *квантове суспільство*; ros. *виртуальная реальность*, *информационное общество*, *цифровой стиль*.



Rysunek 4. Wyniki wyszukiwania dla neologizmu *кіберзлочинність* w PolUkr

Źródło: PolUkr.

[14] Przykłady pochodzą z korpusów równoległych PolUkr i PolRU oraz z internetu.

Podsumowanie

Korpusy równoległe języków słowiańskich ciągle się rozwijają. Ich parametry ilościowe i jakościowe doskonalą się, nadając użytkownikom coraz więcej możliwości. Wykorzystanie korpusów tekstów równoległych oferuje nietrywialne rozwiązania dla wielu kwestii filologicznych. Wśród nich są dwujęzyczna leksykografia i frazeografia, kontrastywne badania dyskursu, semantyka leksykalna, a także teoria przekładu.

W kontekście współczesnych procesów globalizacyjnych rozwój technologii lingwistycznych, umożliwiających skuteczne pokonywanie barier językowych, wychodzi naprzeciw wyzwaniom XXI wieku. Obecnie współczesne narzędzia komunikacji i adaptacji międzyjęzykowej (serwisy typu Tłumacz Google, słowniki internetowe) wciąż są jeszcze niedoskonałe, ale intensywnie się rozwijają, stają się coraz lepsze i coraz dokładniejsze[15]. W ostatnich latach pojawiła się tendencja do budowy dużych infrastruktur naukowych, a także wielojęzycznych zasobów cyfrowych ułatwiających pracę z wielkimi zbiorami tekstów. Wymaga to opracowania i udoskonalenia narzędzi do przetwarzania języka naturalnego, które, biorąc pod uwagę rozwój współczesnej nauki i technologii, prawdopodobnie rozwijać się będą w kierunku dalszej intelektualizacji.

Bibliografia

Blagoeva D. (2008), *Novi frazeologichni kalki v balgarskija ezik (v sapostavka s drugi slavianski ezici)*, [w:] St. Kaldieva-Zaharieva, L. Krumova-Cvetkova (red.), *Izsledvanija po frazeologija, leksikologija i leksikografija*, Sofia.

Bozděchová I. (2010), *Internacionalizáční tendence a typ češtiny*, [w:] *Novye javlenija v slavianskom slovoobrazovanii: sistema i funkcionirovanie*, Moskwa.

Dobrovol'skij D.O. (2009), *Korpus parallelnykh tekstov v issledovanii kulturno-specifichnoj leksiki*, [w:] V.A. Plungian (red.), *Nacionalnyj korpus russkogo jazyka: 2006–2008. Novye rezultaty i perspektivy*, Sankt Petersburg.

Dobrovol'skij D.O. (2015), *Korpusy tekstov i dvujazychnaja frazeografija*, „Vestnik Novosibirskogo gosudarstvennogo pedagogicheskogo universiteta", 5, http://sciforedu.ru/article/1576 [dostęp: 15.04.2019].

Feliksiak M. (2013), *Wulgaryzmy w życiu codziennym*, http://www.cbos.pl/SPISKOM.POL/2007/K_090_07.PDF [dostęp: 15.04.2019].

Hądzlik-Dudka M. (2014), *Wulgaryzmy a przekleństwa w kontekście przemian w komunikacji językowej*, „Studia Filologiczne Uniwersytetu Jana Kochanowskiego", nr 27.

Krztoń W. (2015), *XXI wiek – wiekiem społeczeństwa informacyjnego*, „Modern Management Review MMR", vol. XX(3).

[15] Od 2016 roku w celu zwiększenia płynności i dokładności tłumaczeń serwis internetowy Google Translate wykorzystuje dużą, sztuczną sieć neuronową zdolną do głębokiego uczenia się, https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html [dostęp: 15.04.2019].

Leńko-Szymańska A., Gruszczyńska E. (2016), *Polskojęzyczne korpusy równoległe w Polsce i za granicą*, [w:] E. Gruszczyńska (red.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*, t. I, Warszawa.

Meyer Ch.F. (2004), *English Corpus Linguistics. An Introduction*, Cambridge.

Pędzik P. (2016), *Exploring phraseological equivalence with Paralela*, [w:] E. Gruszczyńska (red.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*, t. I, Warszawa.

Popova T.V. (2005), *Russkaja neologija i neografija*, https://study.urfu.ru/Aid/Publication/174/1/Popova.pdf [dostęp: 15.04.2019].

Satoła-Staśkowiak J. (2016), *Żart w procesie neologizacji – na podstawie języka młodego pokolenia*, „Językoznawstwo", nr 1(10).

Shyrokov V.A. (2017), *Jazyk. Informacija. Sistema. Transdisciplinarnost' v lingvistike*, Saarbrücken.

Sosnowski W., Blagoeva D., Tymoshuk R. (2018), *New Bulgarian, Polish, and Ukrainian phraseology and language corpora*, „Cognitive Studies/Études cognitives", Vol. 18.

Sosnowski W., Tymoshuk R. (2017), *On The dictionary of active Polish and Ukrainian phraseology [Leksykon aktywnej frazeologii polskiej i ukraińskiej]. Contrastive linguistics and culture*, „Cognitive Studies/Études cognitives", Vol. 17.

Styshov O. (2015), *Osnovni dzherela popovnennia frazeolohichnoho skladu ukraïns'koji movy kincia XX – pochatku XXI stolit*', „Movoznavstvo", nr 1, https://movoznavstvo.org.ua/index.php?option=com_attachments&task=download&id=609 [dostęp: 15.04.2019].

Waszakowa K. (2009), *Internacjonalizacja polskiej leksyki – stan obecny, prognozy na najbliższą przyszłość*, [w:] E. Koriakowcewa (red.), *Przejawy internacjonalizacji w językach słowiańskich*, Siedlce, https://ispan.waw.pl/ireteslaw/bitstream/handle/20.500.12528/153/Przejawy%20internacjonalizacji.pdf?sequence=2&isAllowed=y [dostęp: 15.04.2019].

Waszakowa K. (2011), *Polszczyzna przełomu XX i XXI wieku: dynamika procesów sprzyjających internacjonalizacji*, „Issledovanija po Slavianskim Jazykam", nr 16/1.

Źródła internetowe

Clarin-PL, http://clarin-pl.eu/pl/uslugi/ [dostęp: 15.04.2019].

Czeski Korpus Narodowy, https://korpus.cz/ [dostęp: 15.04.2019].

Language Tools and Resources for Polish, http://clip.ipipan.waw.pl/LRT [dostęp: 15.04.2019].

Paralela, http://paralela.clarin-pl.eu/ [dostęp: 15.04.2019].

Seagate, https://www.seagate.com/pl/pl/our-story/data-age-2025/ [dostęp: 15.04.2019].

Słowosieć, http://plwordnet.pwr.wroc.pl/wordnet/ [dostęp: 15.04.2019].

Spokes conversational data search, http://spokes.clarin-pl.eu/ [dostęp: 15.04.2019].

Walenty, http://walenty.ipipan.waw.pl/ [dostęp: 15.04.2019].

Wulgaryzmy w życiu codziennym. Komunikat z badań, https://www.cbos.pl/SPISKOM.POL/2013/K_136_13.PDF [dostęp: 15.04.2019].

Zero-Shot Translation with Google's Multilingual Neural Machine Translation System, https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html [dostęp: 15.04.2019].

Wykaz źródeł

LAFPU – Tymoshuk R., Sosnowski W., Jaskot M., Ganoshenko Y. (2018), *Leksykon aktywnej frazeologii polskiej i ukraińskiej*, Warszawa.

PolRU – *Polsko-rosyjski korpus równoległy*, Clarin-PL, https://clarin-pl.eu/dspace/handle/11321/534 [dostęp: 15.04.2019].

PolUkr – *Polsko-ukraiński korpus równoległy*, Clarin-PL, http://hdl.handle.net/11321/535 [dostęp: 15.04.2019].

WSJP – Żmigrodzki P. (red.), *Wielki słownik języka polskiego*, http://www.wsjp.pl [dostęp: 15.04.2019].

Abstract

Parallel corpora versus language and society, or on the meaning, practical application and prospects for the development of corpus linguistics (illustrated by the example of Polish-Ukrainian and Polish-Russian parallel corpora)

The article discusses issues regarding the role of corpus linguistics and interdisciplinary research in contemporary linguistics. The possibilities of using multilingual digital resources in linguistic research have been discussed and the examples of the use of parallel corpora in research on modern vocabulary and phraseology of Slavic languages have been presented. The considerations lead to the conclusion that nowadays, in times when the need to apply natural language mechanisms in information and computer systems and human-computer interaction is growing, it is necessary to develop resources and language processing tools to effectively overcome language barriers. This will allow for closer cooperation between researchers representing different sciences.

Keywords: language technology, parallel corpus, new vocabulary, phraseology, interlingual equivalence