ROCZNIKI HUMANISTYCZNE
Tom LXVII, zeszyt 7 – 2019
DOI: http://dx.doi.org/10.18290/rh.2019.67.7-5

ANDRZEJ CHARCIAREK

KORPUS RÓWNOLEGŁY INTERCORP W LEKSYKOGRAFII PRZEKŁADOWEJ – MOŻLIWOŚCI I OGRANICZENIA

Polsko-rosyjski słownik par przekładowych [2014] pod redakcją Wojciecha Chlebdy jest chyba najlepszą realizacją na gruncie leksykografii przekładowej polsko-rosyjskiej idei Andrzeja Bogusławskiego, sformułowanej pod koniec lat 80. ubiegłego stulecia. Sprowadzała się ona do wyodrębniania wielowyrazowych jednostek języka, które – jego zdaniem – stanowi jedno z podstawowych zadań współczesnego językoznawstwa. Wspomniany słownik jest wynikiem wyzyskiwania jednostek wielowyrazowych, zwykle zwanych frazemami[1], z istniejących tekstów. Jest to proces niełatwy, ponieważ frazemy, w odróżnieniu od pojedynczych wyrazów, same się graficznie nie wyodrębniają. Z pewnością ta ich cecha powodowała, że tylko pewna część frazemów była notowana w słownikach. Nawet największe słowniki frazeologiczne, zarówno jednojęzyczne, jak i przekładowe, notowały i notują tylko pewną część ich zbioru. Doświadczenia użytkowników słowników, którzy szukają opisu znaczenia czy ekwiwalentu przekładowego określonego frazemu i ich nie znajdują, tylko to potwierdzają. To, dlaczego określone frazemy nie znalazły się

Dr hab. ANDRZEJ CHARCIAREK, prof. UŚ – Zakład Lingwistyki Stosowanej w Instytucie Filologii Wschodniosłowiańskiej UŚ; e-mail: andrzej.charciarek@us.edu.pl

[1] Nazwy *frazem, wielowyrazowa jednostka języka, reprodukt* czy *kolokacja* można traktować jako bliskoznaczne. Cechuje je regularna powtarzalność w tekstach; werbalizując to, co autor chciał powiedzieć w określonej sytuacji – jest to odrębny zespół treściowy (uzewnętrznienie emocji, wypowiedzenie sądu, wyrażenie intencji itp.). W takim ujęciu frazemy tworzą heterogeniczny zbiór jednostek, charakteryzujących się odmiennymi cechami strukturalnymi i semantycznymi.

w tych czy innych opracowaniach leksykograficznych, nie musiało wynikać z przeoczeń czy niekonsekwencji ich autorów, ale z przyjętych przez nich kryteriów wyodrębniania. Przyjęcie np. za podstawę ich notowania w słownikach stanowiska semantycznego autorstwa Stanisława Skorupki powodowało, iż wiele jednostek wielowyrazowych nie mogło być uznanych za frazemy i praktycznie nie były one notowane w słownikach frazeologicznych.

Przyjęcie innych założeń metodologicznych, a przede wszystkim wskazanie innych cech jako podstawowe, diametralnie zmieniło perspektywę widzenia frazemu. Uznanie odtwarzalności i metaforyczności[2] za ich cechy podstawowe spowodowało, że liczba frazemów ogromnie wzrosła, potwierdzając w pewien sposób sąd Andrzeja Bogusławskiego, iż frazemy mogą iść w miliony. Wskazanie cechy reprodukowalności jako podstawowej pozwoliło wskazać frazemy niemające cech typowych jednostek frazeologicznych z ich metaforycznością, obrazowością czy ekspresywnością. Właśnie brak tych właściwości skazywał je wcześniej na status jednostek nieobecnych w słownikach frazeologicznych. Nie było tam miejsca dla frazemów typu *to osobny rozdział*, *ładna historia!*, *i co tam?*, *abyś wiedział!, no to ładnie.* Rodzimy użytkownik języka polskiego natychmiast potwierdzi, że tak się mówi, i co każdy z tych frazemów znaczy; Rosjanin czy Czech, nawet znający dobrze polszczyznę, niekoniecznie. Sytuację można byłoby odwrócić i byłaby ona podobna.

Potrzebę notowania frazemów odczuwa każdy, kto zajmuje się tłumaczeniem. Słowniki przekładowe, w tym wielkie, po które zwykle sięgamy, szukając nawet nie tyle gotowych rozwiązań translatorskich, ile praktycznych wskazówek odnośnie do wykonywanego tłumaczenia, okazują się najczęściej bezradne – nie notują poszukiwanych frazemów i ich ekwiwalentów.

Wspomniany na samym wstępie *Polsko-rosyjski słownik par przekładowych* jest w obszarze leksykografii (nie tylko polsko-rosyjskiej) czymś wyjątkowym i niezwykle pomocnym w procesie przekładu polsko-rosyjskiego, ale przecież i w nim nie znajdziemy rozstrzygnięć wszystkich wątpliwości tłumacza. Każdy słownik przekładowy notuje bowiem ekwiwalenty słownikowe, które są wynikiem pewnego uogólnienia znaczenia frazemu w określonych kontekstach. Nie przypadkiem frazem podany jako ekwiwalent jest w dużym stopniu zdekontekstualizowany, czyli będący w pewien sposób wypadkową kontekstów, w których jest używany. Można zatem traktować go jako jednostkę

[2] Wojciech Chlebda uznaje odtwarzalność za kryterium wydrębniania frazemów (wszystkich), metaforyczność (asumaryczność znaczeniową) natomiast za ich cechę (nie wszystkich). Z tego wynika, że cechą obligatoryjną frazemu jest jego odtwarzalność, natomiast metaforyczność pozostaje cechą fakultatywną.

uniwersalną, która sprawdzi się w wielu tłumaczeniach, ale nie we wszystkich okaże się adekwatna. Przed leksykografem, autorem słownika przekładowego, stoi zatem zadanie dokonania dwóch operacji: znalezienia frazemu w możliwie jak największej ilości kontekstów, a następnie jego zdekontekstualizowania.

W niniejszej pracy skupimy się na tym drugim zabiegu, uznając go za istotniejszy. W tradycyjnej leksykografii przekładowej, czyli kartotekowej, niekomputerowej, na nośniku papierowym, której zmierzch obserwujemy, zadanie leksykografa sprowadzało się do ręcznego gromadzenia frazemów i określenia kontekstów ich użycia. Ponadto leksykograf opracowywał słownik z towarzyszącą mu świadomością ograniczeń natury technicznej – w formie drukowanej nie mógł on przekroczyć określonej objętości.

Bazą materiałową słowników były i są różne źródła, niekiedy niewskazywane przez ich autorów. Dla przykładu zaprezentuję źródła do wydanego w 2002 roku *Słownika frazeologicznego współczesnej polszczyzny* pod redakcją Stanisława Bąby i Jarosława Liberka. Ten niezwykle starannie opracowany słownik jednojęzyczny, rejestrujący polszczyznę końca XX wieku, obejmuje teksty z lat 1968-2000. Wśród nich znajdziemy literaturę piękną (m.in. utwory Zbigniewa Herberta, Stanisława Grochowiaka, Sławomira Mrożka), teksty naukowe i popularnonaukowe z zakresu humanistyki, teksty publicystyczne, prasę (dzienniki, tygodniki i miesięczniki). Jak widać, podstawę ekscerpcji stanowią teksty pisane, co już z założenia stawia poza obszarem badań mówioną odmianę języka, tak przecież bogatą we frazemy. Znamy więc już na samym początku odpowiedź na pytanie, dlaczego te czy inne frazemy nie są w słowniku notowane. Przywołałem ten cenny jednojęzyczny słownik frazeologiczny z dwóch powodów. Po pierwsze po to, aby uzmysłowić, jak ważna jest baza materiałowa tworzonego słownika, który powinien uwzględniać wszystkie rodzaje tekstów w ich obu odmianach: pisanej i ustnej. Po drugie, aby podkreślić wagę badań nad leksykografią jednojęzyczną, w decydujący sposób wpływającą na jakość leksykografii przekładowej poprzez tworzenie jego podstawy – artykułów hasłowych. Jeśli warunek pierwszy nie będzie spełniony, będziemy mogli mówić tylko o fragmentaryczności badań leksykograficznych, coś zawsze w nich zostanie pominięte, niedostrzeżone.

Posłużmy się przykładem odmiennym. Projekt elektronicznego *Wielkiego słownika języka polskiego*, pod patronatem Instytutu Języka Polskiego PAN w Krakowie, powstał w roku 2007 i jest wciąż kontynuowany. W założeniu jest to słownik dokumentacyjny – oparty na autentycznej bazie materiałowej, której podstawę stanowi Narodowy Korpus Języka Polskiego (dalej NKJP), a pomocniczo IJP PAN, Internet oraz ekscerpcja własna pracowników IJP

PAN. Słownik uwzględnia wyłącznie jednostki leksykalne, które zostały poświadczone w wymienionych źródłach. NKJP (bo on nas najbardziej interesuje) liczy około 1,5 mld wyrazów, jego podkorpus zrównoważony ma 240 mln wyrazów. Jeśli chodzi o typy tekstów, to szczegółowo są one opisane w odniesieniu do korpusu zrównoważonego, który zawiera: książki (literaturę piękną, literaturę faktu, książki naukowo-dydaktyczne, książki i prasę informacyjno-poradnikową, prasę (gazety, periodyki), teksty pisane (urzędowe, listy), internet (blogi, fora, strony www), teksty mówione (konwersacyjne naturalne i medialne, protokoły sejmowe). Zwraca uwagę udział tekstów języka mówionego w NKJP, których pozyskiwanie, transkrypcja oraz anotacja jest procesem zarówno czasochłonnym, jak i kosztownym. Obecność polszczyzny mówionej w korpusie jest nieodzowna ze względu na wiarygodność wyników przeprowadzanej analizy ilościowej na podstawie danych korpusowych[3].

Jednak nie NKJP znajdzie się w centrum naszej uwagi, będzie nim tytułowy korpus równoległy InterCorp[4], będący częścią Narodowego Korpusu Języka Czeskiego (Český národní korpus), najstarszego korpusu języka słowiańskiego powstałego w 1994 roku. Jest to korpus wielojęzyczny, rejestrujący wraz z czeszczyzną 40 języków. Jest to korpus otwarty (w odróżnieniu np. od NKJP), czyli stale wzbogacany o nowe teksty. Jego najnowsza, 11. wersja z 2018 roku zawiera teksty publicystyczne i wiadomości ze stron internetowych Project Syndicate i VoxEurop, teksty prawne z korpusu Acquis Communautaire, sprawozdania z obrad Parlamentu Europejskiego (z lat 2007-2011) z korpusu Europarl, napisy filmowe z platformy OpenSubtitles oraz przekłady Biblii. Przytoczone źródła nie rejestrują tekstów we wszystkich językach, co oznacza, że dla każdej leksykografii przekładowej z osobna, np. polsko-czeskiej, czesko-rosyjskiej czy polsko-rosyjskiej, InterCorp stwarza odmienne, lepsze lub gorsze warunki.

Warto w tym miejscu podkreślić, że InterCorp służy w pierwszym rzędzie do badania języka czeskiego, a poza nim pozycję uprzywilejowaną ma język angielski. Nic zatem dziwnego, że dla leksykografii czesko-angielskiej InterCorp stanowi nieocenione źródło. Wartość tego korpusu równoległego dla pozostałych leksykografii przekładowych zależy przede wszystkim od wielkości kor-

[3] Przeprowadzenie tej operacji postulował już w 2006 roku Piotr Żmigrodzki, pisząc: „Konieczne wydaje się też powiększenie w korpusach udziału tekstów języka mówionego, zwłaszcza wypowiedzi spontanicznych, reprezentujących odmianę potoczną (nieoficjalną)" [Żmigrodzki 2006, 177].

[4] http://ucnk.korpus.cz/intercorp

pusów poszczególnych języków. Omówmy po kolei korpusy równoległe: polsko-czeski, czesko-rosyjski i polsko-rosyjski.

Polsko-czeski InterCorp jest największym korpusem z wymienionych, a to za sprawą wielkości korpusu polskiego liczącego ponad 86 mln wyrazów. Zawiera on teksty literatury pięknej (ponad 24 mln wyrazów), teksty prawne z korpusu Acquis Communautaire (blisko 20 mln wyrazów), sprawozdania z obrad Parlamentu Europejskiego z korpusu Europarl (prawie 13 mln), teksty publicystyczne i wiadomości ze stron internetowych VoxEurop (prawie 2,5 mln wyrazów), napisy filmowe z bazy OpenSubtitles (około 27 mln wyrazów) oraz przekłady Biblii (ponad 0,5 mln wyrazów). Zwraca uwagę fakt, że w korpusie polsko-czeskim brakuje tylko jednej kolekcji tekstów w InterCorp, czyli publicystyki w języku polskim ze strony Project Syndicate[5].

Wyraźnie mniejszy rozmiar ma rosyjski InterCorp (ponad 18 mln wyrazów), który oprócz tradycyjnego rdzenia (czes. *jadro*), czyli tekstów beletrystycznych (ponad 7 mln wyrazów), zawiera teksty publicystyczne ze strony Project Syndicate (niespełna 4 mln wyrazów, napisy filmowe z bazy OpenSubtitles (prawie 7 mln wyrazów) i przekłady Biblii (ponad 0,5 mln wyrazów).

Niewielkie rozmiary korpusu tekstów w języku rosyjskim mają konsekwencje odnośnie do zestawianych z nimi tekstów czeskich i polskich. Rosyjsko-czeski InterCorp zawiera wszystkie wymienione 4 kolekcje w języku rosyjskim i ich czeskie odpowiedniki. Rosyjsko-polski InterCorp jest uboższy o teksty publicystyczne ze strony Project Syndicate, zawiera beletrystykę, napisy filmowe i przekłady Biblii.

Z powyższych charakterystyk jasno wynika, że w zależności od zestawianych ze sobą języków, określony korpus równoległy może być mniej lub bardziej przydatny w określonej leksykografii przekładowej. Wszystko to za sprawą typów i liczby tekstów, które się w nim znajdują.

Pojęcie *korpus równoległy* jest często odnoszone, i nie bez przyczyny, do korpusów dwujęzycznych, w których zestawione są ze sobą oryginał i jego przekład. Takich korpusów znajdziemy wiele. Jako przykład może posłużyć Polsko-Rosyjski i Rosyjsko-Polski Korpus Równoległy Uniwersytetu Warszawskiego, zawierający głównie oryginały w języku polskim lub rosyjskim i ich przekłady[6]. InterCorp jest natomiast korpusem równoległym wielojęzycznym, w którym każdy tekst obcojęzyczny ma odpowiednik w języku czeskim.

[5] Teksty publikowane na stronach Project Syndicate tłumaczone są na język arabski, chiński, czeski, francuski, holenderski, hiszpański, portugalski, rosyjski i szwedzki.

[6] Korpus ten zawiera również kilka tekstów literackich w języku angielskim, niemieckim i francuskim oraz ich przekłady polskie i rosyjskie.

Innymi słowy, tekst w języku czeskim może być albo oryginałem, albo przekładem. Zestawianie zatem ze sobą tekstów równoległych w dwu językach może być zestawianiem nie oryginału z jego przekładem, ale przekładów. W praktyce, zresztą częstej, choć oryginałami bywają teksty w języku angielskim, to zestawiane są ze sobą np. jego polskie i czeskie przekłady. To, z jakim materiałem językowym ma do czynienia użytkownik korpusu InterCorp, jest wiadomo dzięki szczegółowej anotacji zewnętrznej tekstów. W wypadku literatury pięknej informacje dotyczą: tytułu, autora, tłumacza, języka oryginału, języka przekładu, roku i miejsca wydania, wydawnictwa. Nawet w wypadku nieprofesjonalnych napisów filmowych podawana jest informacja, o tekście źródłowym i kierunku tłumaczenia.

Powyższe informacje są istotne, bowiem użycie określonych frazemów może być wynikiem ich nadużywania zarówno przez autora, jak i tłumacza tekstu. W takiej sytuacji odpowiednia anotacja korpusowa pozwoli na wychwycenie nawet sporej liczby poświadczeń, których autorem lub tłumaczem jest jedna i ta sama osoba. Są to ważne informacje dla leksykografa pozyskującego poświadczenia tych czy innych frazemów.

W przypadku napisów filmowych z platformy OpenSubtitles dane korpusowe wymagają szczególnej weryfikacji, ponieważ tłumaczenia mają charakter amatorski, a ich autorzy z reguły nie posiadają odpowiednich kwalifikacji i kompetencji językowych. Tak więc ryzyko pozyskania poświadczeń z błędami tłumaczeniowymi i językowymi w przypadku tych zasobów korpusowych jest spore. Trudno porównywać je z napisami profesjonalnymi, wykonywanymi przez specjalistów według ściśle określonych kryteriów oraz z zastosowaniem odpowiedniego sprzętu i specjalistycznego oprogramowania.

Jak wiadomo, profesjonalne napisy filmowe powstają na podstawie listy dialogowej, a ta wynika ze scenariusza. To właśnie tworzenie napisów jest procesem, który w znacznym stopniu modyfikuje (przede wszystkim kondensuje) wypowiedzi postaci filmu. Przyczyną tego zjawiska są ograniczenia natury technicznej, czyli czasowe, przestrzenne, sekwencyjne i graficzne. Każde z nich można traktować jako ingerencję w tekst oryginału, a następnie tłumaczenia. Nie należy zapominać także o tym, że w wypadku tworzenia napisów filmowych mamy do czynienia z tłumaczeniem intralingwalnym – oryginał, który jest tekstem mówionym, zmienia swoją postać na tekst pisany. Ten ostatni stanowi podstawę tłumaczenia na język docelowy, podlegający również adaptacji do wymogów technicznych.

Amatorskie opracowywanie napisów filmowych przebiega zwykle całkiem inaczej. Nierzadko tłumacz amator nie stosuje się do zasad sporządzania napi-

sów filmowych, nie dysponuje także wspomnianym zapleczem technicznym, a przede wszystkim listą dialogową filmu. Napisy tworzy najczęściej na podstawie odsłuchu ścieżki dźwiękowej filmu, co może być kolejnym źródłem potknięć. Na końcowy efekt napisów filmowych z platformy OpenSubtitles składa się więc wiele czynników. Nic zatem dziwnego, że zestawiane w korpusie InterCorp napisy (oryginał i przekład lub przekład i przekład) mają ze sobą niekiedy niewiele wspólnego. Jednak ten ze wszech miar niedoskonały materiał korpusowy posiada również zalety. Choć nie można traktować napisów filmowych jako zapis *sensu stricto* języka mówionego, to jednak wykazują z nim wiele podobieństw – zawierają wypowiedzi spontaniczne, potoczne w sytuacji nieoficjalnej. Właśnie one są praktycznie nieobecne w innych rodzajach tekstów i dlatego brakuje ich poświadczeń korpusowych. Biorąc pod uwagę stale poszerzającą się sferę potoczności, tym bardziej wzrasta wartość rejestracji w korpusie zachowań werbalnych, wykorzystywanych w sytuacjach codziennej, nieoficjalnej komunikacji. Znajdziemy zatem w napisach filmowych frazemy o dużym stopniu idiomatyczności, oczywiste dla rodzimych użytkowników określonego języka, ale już nie dla nierodzimych. Jak wiemy, próby znalezienia ekwiwalentów idiomów w języku obcym *per analogiam* do języka rodzimego czy, odwrotnie, przynoszą zwykle marny skutek. Oprócz tego to, co wydawać się może ewidentną wadą napisów z bazy OpenSubtitles, czyli łamanie zasad w procesie ich tworzenia, paradoksalnie może stać się ich zaletą. Dobrym przykładem jest zasada wierności oryginałowi, często podkreślana przez tłumaczy amatorów, a sprowadzająca się do osiągnięcia adekwatności przekładu nawet za cenę nierespektowania ograniczeń technicznych, zwłaszcza czasowo-przestrzennych. Cała przytoczona charakterystyka amatorskich napisów filmowych pokazuje, z jak zróżnicowanym materiałem językowym użytkownik korpusu równoległego może mieć do czynienia. Wielką zaletą tego materiału jest rejestracja jednostek, których często próżno szukać w słownikach przekładowych.

Należy wspomnieć o tym, że notowanie języka mówionego w korpusie InterCorp nie ogranicza się tylko do napisów filmowych. Stylizację języka mówionego spotkamy w dialogach utworów literatury pięknej, które notowane są w każdym z trzech analizowanych korpusów (czeskim, polskim i rosyjskim).

W wypadku polsko-czeskiego korpusu InterCorp odrębną kolekcję stanowią sprawozdania z obrad Parlamentu Europejskiego, czyli wypowiedzi w sytuacji oficjalnej. Użytkownik znajdzie zatem w tym materiale frazemy przynależne do stylu administracyjnego, publicystycznego czy naukowego. Jego

uzupełnieniem będą frazemy z tekstów prawnych z korpusu Acquis Communautaire.

Brak wymienionych dwu kolekcji tekstów w korpusie rosyjskim znacznie obniża jego wartość leksykograficzną. W niewielkim stopniu ten brak rekompensują teksty publicystyczne ze stron Project Syndicate, które stwarzają dodatkowe możliwości dla tłumaczenia czesko-rosyjskiego.

W niniejszym tekście często podnoszę wagę notowania w korpusie równoległym zapisów języka mówionego, ponieważ jest to główne źródło jednostek tworzących idiomatykę komunikacyjną, rzadko notowaną w słownikach. Nie wydaje się, aby sytuacja z niewystarczającą ilością zapisów języka mówionego w korpusach równoległych radykalnie się zmieniła. Po pierwsze, ich pozyskiwanie jest czasochłonne i kosztowne, a po drugie, nie ma obiektywnych przesłanek, które sprzyjałyby tłumaczeniu tego typu tekstów na inne języki; oczywiście, poza potrzebami wzbogacania i powiększania korpusów równoległych czy rozwojem leksykografii przekładowej.

Niepodobna pominąć w omówieniu przydatności korpusu InterCorp w leksykografii przekładowej zagadnienia kluczowego, jakim jest ekwiwalencja przekładowa. I choć doczekała się ona wielu opracowań, to wciąż pozostaje w centrum uwagi wielu badaczy.

Chyba najbardziej syntetyczne omówienie ekwiwalencji przekładowej, rozumianej jako równoważność komunikacyjna oryginału i przekładu, znajdziemy w monografii *Zagadnienia lingwistyki przekładu* Romana Lewickiego [Lewicki 2017]. Autor wskazuje w niej sześć cech ekwiwalencji przekładowej:

1. Ma charakter asymetryczny. Jeden z tekstów jest oryginałem, drugi – jego przekładem. Odwrócenie tego układu nie jest możliwe.

2. Jest cechą przekładu i polega na podobieństwie przekładu do oryginału poprzez jego naśladowanie.

3. Jest gradualna, co oznacza, że przekład jest stopniowalny – może być w mniejszym lub większym stopniu ekwiwalentny wobec oryginału.

4. Jest względna, ponieważ może być oceniana wedle różnych kryteriów.

5. Jest hierarchiczna, ponieważ podlega jej nie tylko przekład jako całość, ale także jego części.

6. Nie jest oparta na zasadzie równoznaczności wyrażeń językowych użytych w oryginale i w jego przekładzie. Jej celem jest uzyskanie przekładu równoważnego w stosunku do oryginału. Innymi słowy, przekład powinien posiadać analogiczną wartość komunikacyjną co oryginał [Lewicki, 138-139].

Odniesienie wszystkich cech ekwiwalencji przekładowej do materiału zgromadzonego w korpusie równoległym nie jest możliwe. Przede wszystkim,

jak już wspomniałem, często mamy do czynienia z innym układem relacji niż typowy, a mianowicie: oryginał – przekład. Dysponujemy zatem głównie przekładami[7] (często z języków, których nie znamy w ogóle lub znamy słabo), które w różnym stopniu naśladują oryginał, w większym czy mniejszym stopniu są wobec niego ekwiwalentne. Jednak nie można twierdzić, że nie ma związku między przekładami. Translaty (jednostki języka przekładu), np. polski i czeski, są determinowane przez ten sam obcojęzyczny transland (jednostkę języka źródłowego), który – co by nie powiedzieć – w pewien sposób uczestniczy w procesie ustanawiania pary przekładowej[8]. Jego obecność – co oczywiste – jest zjawiskiem niepożądanym, utrudniającym nierzadko określenie relacji między właściwym translandem (np. polskim) a translatem (np. czeskim). Niemniej rezygnacja z przekładów z innych języków i badanie wyłącznie tłumaczeń bezpośrednich powoduje drastyczne zmniejszenie liczby dokumentów, a co za tym idzie – także liczby poświadczeń. Należy pamiętać, że zbyt mała liczba poświadczeń może uniemożliwić przeprowadzenie zobiektywizowanych badań nad określonym zjawiskiem językowym[9]. Możemy wobec tego mówić o pewnym kompromisie badawczym, polegającym na poszerzeniu materiału korpusowego przy jednoczesnym obniżeniu jego jakości. Skutkować to będzie nieraz – o czym już wspomniałem – różnego rodzaju nieporozumieniami w określaniu ekwiwalentów, a tym samym nieadekwatnością wyodrębnianych par przekładowych. Wspomniane nieporozumienia czy pomyłki mogą być dodatkowo spotęgowane przez takie czy inne techniczne opracowanie materiału korpusowego. Mam tu na myśli głównie ręczne lub automatyczne wyrównanie tekstów (ang. *alignment*), dokonywane zwykle na poziomie zdań. W korpusie InterCorp tylko teksty literatury pięknej są opracowane ręcznie, wszystkie pozostałe kolekcje (Acquis, VOX, Europarl, Open Subtitles, Project Syndicate, Biblia) – automatycznie. Choć oprogramowanie do automatycznego opracowania tekstów jest konsekwentnie doskonalone, to

[7] Korpus równoległy InterCorp zawiera głównie przekłady, przy czym każdy tekst obcojęzyczny ma swój czeski odpowiednik (oryginał lub przekład). Nie należy zapominać o tym, że InterCorp służy głównie do badania języka czeskiego w porównaniu z innymi językami.

[8] W niniejszej pracy jako translandy traktuję frazemy (jednostki wielowyrazowe) języka źródłowego, natomiast jako translaty – frazemy języka przekładu. To, oczywiście, nie wyklucza, że translandami czy translatami mogą być inne jednostki tekstu – zarówno mniejsze (np. wyraz), jak i większe (np. zdanie).

[9] Ta oczywista kwestia podnoszona jest w wielu pracach z zakresu lingwistyki korpusowej, np. przez Milenę Hebal-Jezierską, Alexandra Rosena, i Elżbietę Kaczmarską w artykule *Between the devil and the deep blue sea or between users' needs and the compilers' powers: An analysis of the Czech-Polish part of the parallel corpus InterCorp* [Warszawa 2016].

ustępuje ono metodzie ręcznej. Wprawdzie wyrównanie na poziomie zdań nie powoduje komplikacji w sytuacji, gdy jedno zdanie w jednym języku (np. w oryginale) odpowiada jednemu zdaniu w drugim języku (np. w przekładzie), to inaczej jest, jeśli stosunek jeden do jednego (jedno zdanie do jednego zdania) zastąpiony zostanie innymi kombinacjami, np. 1:2, 1:3 czy 2:1. Z takimi sytuacjami opracowanie automatyczne nie zawsze może sobie poradzić. Nierzadko więc właściwy ekwiwalent przekładowy znajduje się poza fragmentem prezentowanym na ekranie. Jego znalezienie jest możliwe po rozszerzeniu kontekstu.

Wszystkie wyżej opisane czynniki mogą wpływać na uzyskiwane przez nas wyniki analizy korpusowej. Jeśli w stosunku do wyników analizy kwantytatywnej, przeprowadzanej w korpusie jednojęzycznym, należy zachować ostrożność, to w przypadku korpusu równoległego (tu – InterCorp) ta ostrożność powinna być jeszcze większa[10]. Ryzyko wyciągania nieuprawnionych, pochopnych wniosków jest także znacznie większe. Wskazanie optymalnego ekwiwalentu przekładowego nie musi być wcale wynikiem większej liczby jego poświadczeń. Okazuje się nierzadko, że translat, który posiada mniej poświadczeń w korpusie, jest w większym stopniu ekwiwalentny niż pozostałe częściej notowane[11]. InterCorp nie podaje gotowych rozwiązań przekładowych, prezentuje jedynie możliwości tłumaczenia określonych fragmentów tekstu. Można nawet stwierdzić, że naprowadza na odpowiedni wariant tłumaczenia. Użytkownik korpusu równoległego widzi próby naśladowania oryginału czynione poprzez poszczególnych tłumaczy, reprezentujących przecież odmienne idiolekty. Ich wybory translatorskie są zdeterminowane przez czynniki socjodemograficzne: wiek, płeć, wykształcenie itd. To dlatego wyniki analizy materiału w korpusie równoległym często zaskakują badaczy, ponieważ nie odpowiadają ich indywidualnym, stereotypowym oczekiwaniom.

W wypadku frazemów jest to bardzo istotne, ponieważ, jak wiadomo, najczęściej nie posiadają one jednego ekwiwalentu przekładowego. Szczególnie chodzi tu o frazemy, które mają znaczenie metaforyczne, obrazowe, niebędące sumą znaczeń elementów, które zawierają. To właśnie InterCorp oferuje zbiór ich możliwych odpowiedników przekładowych, które, w zależności od sytu-

[10] Tadeusz Piotrowski i Łukasz Grabowski piszą o zasadzie ograniczonego zaufania w odniesieniu do wyników pozyskiwanych ze współczesnych jednojęzycznych korpusów lingwistycznych [Piotrowski, Grabowski 2013, 65].

[11] Z moich analiz nieraz wynikało, że odpowiednik przekładowy, najbardziej ekwiwalentny wobec oryginału, miał mniej poświadczeń niż inne mniej ekwiwalentne.

acji ich użycia, są w mniejszym czy większym stopniu ekwiwalentne wobec oryginału.

Analiza kwantytatywna wydatnie pomaga w ustaleniu pary przekładowej, ale nie stanowi wyłącznego kryterium. Z pewnością przynosi ona ważne informacje statystyczne, które powinny być uwzględnione przez leksykografa. Informacje te są ważnym argumentem w procesie ustalania translatów w języku docelowym, pozwalającym rozstrzygnąć różne dylematy leksykografów co do kolejności notowania możliwych ekwiwalentów w słowniku przekładowym. Te ostatnie tworzą przecież ciągi frazemów synonimicznych, o których kolejności mogą decydować różne kryteria, w tym kryterium statystyczne. Frekwencja dokumentuje cechę podstawową frazemów – stopień ich reprodukowalności. Wykorzystanie korpusu równoległego pozwala na wychwycenie stopnia powtarzalności określonych frazemów i ich dystrybucji w poszczególnych rodzajach tekstów. Nie bez znaczenia pozostaje możliwość określenia za pomocą tekstów korpusowych wariantów leksykalnych frazemu, który uznawany jest za kanoniczny. Innymi słowy, InterCorp może posłużyć jako narzędzie służące do wyodrębniania wariantów frazemów i ich szeregowania, jeśli weźmie się pod uwagę dane statystyczne[12].

Powyższe rozważania są rezultatem przeprowadzonych przeze mnie badań korpusowych nad wybranymi frazemami (translandami) pod kątem wskazania ich obcojęzycznych translatów[13]. Za każdym razem bazę materiałową i zarazem narzędzie badawcze stanowił korpus równoległy InterCorp. W zależności od zestawianych języków układy przybierały różne kierunki tłumaczenia: czeski – polski, polski – czeski, polski – rosyjski, rosyjski – polski. Niekiedy analizę wspierały jednojęzyczne korpusy narodowe, poniekąd potwierdzając tezę, że korpus InterCorp nie jest pozbawiony wad.

Poszczególne analizy jednoznacznie potwierdziły przydatność korpusu równoległego w poszukiwaniu i ustalaniu ekwiwalentów przekładowych. W odróżnieniu od wyników analizy materiału korpusowego, artykuły hasłowe frazemów w tradycyjnych, papierowych słownikach przekładowych okazywały się niepełne, rejestrujące tylko niektóre z możliwych ekwiwalentów. Nie mówiąc o tym, że niektóre translaty, mające wysoką frekwencję użycia, bywały często pomijane, a ich miejsce zajmowały inne, praktycznie nieużywane. Fakt ten nie wynika – co oczywiste – z braku profesjonalizmu autorów słowników, ale

[12] Wspomina o tym Piotr Żmigrodzki w odniesieniu do jednojęzycznych słowników elektronicznych [Żmigrodzki 2009, 33].

[13] Zostały one opublikowane w Brnie, Gdańsku, Pradze i Warszawie. Wszystkie przywołuję w bibliografii.

często nietrafnie określonej bazy materiałowej poddanej analizie[14]. W odróżnieniu od słownika, korpus InterCorp okazywał się zdecydowanie bliższy temu, co możemy nazwać oglądem rzeczywistego użycia języka.

Wspomniana przydatność korpusu InterCorp, wraz z powiększaniem się jego zasobów, będzie – o czym jestem przekonany – sukcesywnie wzrastać. Można również pokusić się o konkluzję natury ogólniejszej: dane empiryczne, a takimi przecież są dane korpusowe, będą miały coraz większy wpływ na badania nad językiem traktowanym jako tekst, w którym dychotomia centrum – peryferie przybiera formę: jednostki często używane – jednostki rzadko używane[15].

BIBLIOGRAFIA

Charciarek, Andrzej. "Możnosti využití korpusu InterCorp v česko-polské překladové lexikografii". Časopis pro moderní filologii 100.2 (2018): 206-222.

Charciarek, Andrzej. "Korpus równoległy InterCorp w leksykografii przekładowej polsko-rosyjskiej". Słowo z perspektywy językoznawcy i tłumacza. Vol. 7. Frazeologia z perspektywy językoznawcy i tłumacza. Eds. Pstyga, Alicja, Tatiana Kananowicz, and Magdalena Buchowska. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego, 2018. 54-66.

Charciarek, Andrzej. "Параллельный корпус как инструмент польско-русской переводной лексикографии" [Parallel'nyy korpus InterCorp kak instrument pol'sko-russkoy perevodnoy leksikografii]. Język rosyjski XXI wieku. Żródła i perspektywy. Warszawa: Instytut Rusycystyki Uniwersytetu Warszawskiego, 2017. 151-164.

Charciarek, Andrzej. "Параллельный корпус InterCorp в переводной лексикографии" [Parallel'nyy korpus InterCorp v perevodnoy leksikografii]. Opera Slavica 2 (2017): 5-17.

Čvrček, Vaclav, et al. Mluvnice současné češtiny 1. Jak se píše a jak se mluví. Praha: Nakladatelství Karolinum, 2015.

Hebal-Jezierska, Milena, Aleksander Rosen, and Elżbieta Kaczmarska. "Between the devil and the deep blue sea or between users' needs and the compilers' powers: An analysis of the Czech-Polish part of the parallel corpus InterCorp". Polskojęzyczne korpusy równoległe. Eds. Gruszczyńska, Ewa, and Agnieszka Leńko-Szymańska. Warszawa: Wydział Lingwistyki Stosowanej Uniwersytetu Warszawskiego, 2016. 41-66.

Lewicki, Roman. Zagadnienia lingwistyki przekładu. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, 2017.

[14] Bolączką wielu opracowań leksykograficznych, zwłaszcza o charakterze przekładowym, jest niekiedy bezkrytyczne zaufanie do wcześniej wydanych słowników, a co za tym idzie, częstokroć powielanie zawartych w nich błędów.

[15] Najlepszym tego przykładem jest, co prawda, nie słownik, ale *Gramatyka współczesnego języka czeskiego* opracowana na materiale językowym z Narodowego Korpusu Języka Czeskiego: V. ČVRČEK ET AL., *Mluvnice současné češtiny*, 1: *Jak se píše a jak se mluví*, Praha: Nakladatelství Karolinum 2015.

Piotrowski, Tadeusz, and Łukasz Grabowski. ”Interpretacja danych frekwencyjnych z korpusów językowych: opis pewnych problemów (na kilku przykładach z życia wziętych)”. Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów. Ed. Wojciech Chlebda. Opole: Wydawnictwo Uniwersytetu Opolskiego, 2013. 59-71.

Żmigrodzki, Piotr. Wprowadzenie do leksykografii polskiej. Katowice: Wydawnictwo Uniwersytetu Śląskiego, 2009.

Żmigrodzki, Piotr, Renata Przybylska, and Dunaj Bogusław. ”O potrzebie nowego słownika języka polskiego”. LingVaria” 1 (2006): 171-179.

SŁOWNIKI

Polsko-rosyjski słownik par przekładowych. Tom zbiorczy podręcznego idiomatykonu polsko-rosyjskiego (z. 1-5). Ed. Chlebda, Wojciech. Opole: Wydawnictwo Uniwersytetu Opolskiego, 2014.

Bąba, Stanisław, and Jarosław Liberek. Słownik frazeologiczny współczesnej polszczyzny. Warszawa: Wydawnictwo Naukowe PWN, 2002.

KORPUS RÓWNOLEGŁY INTERCORP W LEKSYKOGRAFII PRZEKŁADOWEJ – MOŻLIWOŚCI I OGRANICZENIA

S t r e s z c z e n i e

Niniejszy artykuł poświęcony jest teoretycznej refleksji nad wykorzystaniem korpusu równoległego InterCorp w leksykografii przekładowej. Opisano zasoby tekstowe i specyfikę poszczególnych modułów językowych korpusów równoległych: polskiego, czeskiego i rosyjskiego. Wskazano zarówno zalety, jak i wady poszczególnych dwujęzycznych korpusów równoległych: polsko-czeskiego, polsko-rosyjskiego i czesko-rosyjskiego. Wśród kwestii teoretycznych skupiono się głównie na zagadnieniu ekwiwalencji przekładowej i jej kryteriach w odniesieniu do materiału korpusowego, zawierającego w większości przekłady. Wykazano wciąż wzrastającą przydatność korpusu równoległego InterCorp w leksykografii przekładowej.

Słowa kluczowe: leksykografia przekładowa; korpus równoległy InterCorp**;** ekwiwalencja przekładowa**;** frazemy; język czeski; język polski.

INTERCORP PARALLEL CORPUS IN TRANSLATION LEXICOGRAPHY – OPPORTUNITIES AND LIMITATIONS

Summary

The article is devoted to theoretical considerations related to the use of InterCorp parallel corpus in the translation lexicography. It provides a description of text resources and specificity of particular linguistic models of paralel corpora: Polish, Czech and Russian. As well as this, advantages and disadvantages of bilingual paralel corpora (Polish-Czech, Polish-Russian and Russian-Czech) are discussed. Theoretical issues focus mainly on translation equivalence and its criteria in reference to the corpus resources including mostly translations. The study proves increasing usability of the InterCorp parallel corpus in translation lexicography.

Key words: translation lexicography; Intercorp parallel corpus; translation equivalence; phrasemes; Czech; Polish.