Elżbieta Kaczmarska
Uniwersytet Warszawski, Warszawa
E-mail: e.h.kaczmarska@uw.edu.pl
ORCID: 0000-0002-8838-1404

https://doi.org/10.18778/8220-635-7.12

INTERCORP I TREQ JAKO NARZĘDZIA UŁATWIAJĄCE WYSZUKIWANIE EKWIWALENTÓW

Niemożność oddania w języku docelowym dokładnie tego samego, co wyrażone jest w języku wyjściowym, często wiąże się z wieloznacznością i problemem precyzyjnego zrozumienia przekładanej jednostki (Kaczmarska, 2019; Kaczmarska & Rosen, 2014a). Szczególnie kłopotliwymi w tym kontekście jawią się czasowniki dotyczące różnych stanów psychicznych czy emocjonalnych, które już same w sobie mogą być źródłem niejasności i nieporozumień (Kaczmarska, 2016, 2019). Czasowniki te sprawiają wiele trudności przy przekładzie; często nie jest jasne, jak je przełożyć i nie mniej często pada pytanie, co właściwie znaczą. Jednostki te są na ogół wieloznaczne i poszukując ekwiwalentów, napotyka się całe ich łańcuchy, nie można jednak znaleźć jednego konkretnego, który by w pełni oddawał znaczenie, styl, treść oryginału (Lewandowska-Tomaszczyk, 1984, 2013). W związku z tym część treści zostaje utracona w przekładzie; tłumaczenie jest uboższe i w efekcie miewa zmodyfikowane znaczenie. Pomocą powinny służyć w tym miejscu słowniki, jednakże słowniki klasyczne (papierowe) ze względu na swoją objętość nie mogą zawrzeć wszystkich znaczeń wyrazów wieloznacznych wraz z kontekstami i przykładami, a te byłyby w takich przypadkach konieczne. Pomocne mogą być narzędzia do ujednoznaczniania znaczeń; dla języka polskiego – WoSeDon[1] czy Słowosieć[2]. W przyszłości obydwa

[1] WoSeDon to narzędzie przeznaczone do ujednoznaczniania znaczeń słów (ang. *Word Sense Disambiguation*). Zostało zaprojektowane dla języka polskiego, jednak po odpowiedniej zmianie tagsetu, możliwe byłoby działanie również na innych językach. Głównym algorytmem wykorzystanym w tym narzędziu, który został zaimplementowany do rozstrzygania niejednoznaczności, jest PageRank. WoSeDon zawiera różne modyfikacje tego algorytmu oraz umożliwia manipulacje wieloma parametrami, które mają wpływ na jakość ujednoznaczniania znaczeń leksykalnych (Janz i in., 2018; Kaczmarska, 2019; Kędzia i in., 2016).

[2] Słowosieć (z ang. *wordnet*) to relacyjny słownik semantyczny, który odzwierciedla system leksykalny języka polskiego (Dziob & Piasecki, 2018; Maziarz i in., 2014; Piasecki i in., 2016). Pojedyncze znaczenia w Słowosieci połączone są wzajemnymi relacjami. Tak powstaje sieć, w której każdy wyraz jest zdefiniowany poprzez odniesienie do innych wyrazów, np. *zając* to zwierzę; *zając* stanowi całość, na którą składają się np. głowa, słuchy, skoki, omyk, trzeszcze, turzyca, natomiast jego wyrazami bliskoznacznymi są szarak, marczak, gach, defilator. Słowosieć to także bardzo ważny zasób, biorący

narzędzia mogłyby być wykorzystywane w procesie ujednoznaczniania znaczeń podczas badań dwujęzycznych (Kaczmarska, 2019). Obecnie rolę tę mogą spełniać częściowo korpusy równoległe; w przypadku badań i przekładów czesko-polskich pomocą może służyć korpus paralelny InterCorp[3] (Čermák & Rosen, 2012) i oparta na nim baza ekwiwalentów kontekstowych Treq[4]. Przy użyciu tych narzędzi[5] w niniejszym artykule analizowane są możliwości ujednoznacznienia znaczenia czasownika *mrzet* oraz znalezienia dla niego ekwiwalentu w języku polskim.

InterCorp

InterCorp[6] (Čermák & Rosen, 2012; Kaczmarska, 2019; Kaczmarska & Rosen, 2014b; Rosen, 2016; Rosen & Vavřín, 2014; Rosen i in., 2019) to akademicki i niekomercyjny projekt, który powstał w Pradze na Wydziale Filozoficznym Uniwersytetu Karola. Jego celem jest wybudowanie obszernego równoległego korpusu synchronicznego, który zawierałby teksty z jak największej liczby języków[7]. InterCorp jest jednocześnie nazwą tego ciągle rozrastającego się synchronicznego korpusu paralelnego, obejmującego aktualnie 41 języków (stan na maj 2021 roku). InterCorp jako korpus jest też częścią Czeskiego Korpusu Narodowego[8].

InterCorp zawiera szereg jednobrzmiących tekstów, których „parą" jest zawsze tekst czeski (bądź jako oryginał, bądź jako tłumaczenie). Język czeski jest więc dla InterCorpu językiem kluczowym[9].

---

udział w komputerowym przetwarzaniu języka i badaniach nad sztuczną inteligencją, znajdującym m.in. zastosowanie w automatycznych tłumaczeniach Google Translate. Polska Słowosieć jest już największym *wordnetem* na świecie i nieustannie się rozrasta. Por. WordNet (Fellbaum, 1998; Miller 1995).

[3] InterCorp oferuje zestawy ekwiwalentów wraz z przykładami i (szerokimi) kontekstami.

[4] Treq automatycznie sporządza listę ekwiwalentów dla danego słowa.

[5] Podrozdziały poświęcone korpusowi InterCorp i bazie Treq powstały na podstawie monografii *Metody ustalania ekwiwalentów czasowników wyrażających stany emocjonalne w przekładzie czesko-polskim na materiale z korpusu równoległego InterCorp* (Kaczmarska, 2019).

[6] Źródło i dostęp: http://www.korpus.cz/intercorp/

[7] W pierwotnym założeniu miały to być wszystkie języki nauczane na Wydziale Filozoficznym Uniwersytetu Karola w Pradze.

[8] www.korpus.cz

[9] Inaczej wygląda sytuacja w przypadku korpusu ParaSol, również wielojęzycznego korpusu równoległego: „W odróżnieniu od znacznie większego korpusu InterCorp w ParaSolu większą wagę przykłada się do reprezentacji i zrównoważenia jak największej liczby języków i żaden z tych języków nie jest podstawowy, tak jak podstawowy jest czeski w korpusie InterCorp, gdzie każdy tekst musi mieć czeską wersję" (Łaziński, 2014, s. 200). Por. Łaziński i in., 2012; von Waldenfels, 2012.

Korpus równoległy[10] służy między innymi jako źródło danych do badań teoretycznych, prac translatorskich, analiz gramatycznych i leksykograficznych, projektów dotyczących nauki języków obcych, aplikacji komputerowych, czy poszukiwań studenckich[11].

Opis korpusu InterCorp

Korpus InterCorp składa się z dwóch części – trzonu (tzw. *core*) i kolekcji (*collections*). Trzon stanowią przede wszystkim teksty beletrystyczne, ręcznie wiązane[12]. Oprócz nich korpus obejmuje również automatycznie opracowane teksty; są to widoczne w interfejsie wspomniane wcześniej kolekcje. Obecnie w ramach tych kolekcji udostępnione są artykuły publicystyczne, teksty ze stron Project Syndicate oraz Presseurop, akty prawne Unii Europejskiej z korpusu Acquis Communautaire, zapisy posiedzeń Parlamentu Europejskiego z lat 2007–2011 (z korpusu Europarl), bazy napisów dialogowych z serweru Open Subtitles, a także *Biblia*. Teksty w kolekcjach są wiązane automatycznie, dlatego w wyszukanych konkordancjach może pojawić się więcej zdań, które sobie nawzajem nie odpowiadają. Wszystkie teksty w InterCorpie mają zawsze wersję czeską (jako oryginał lub tłumaczenie); jak już wcześniej zostało wspomniane, język czeski jest dla InterCorpu językiem kluczowym (*pivot*), a czeska wersja jest wiązana z jedną lub wieloma wersjami innojęzycznymi.

Najnowszą edycją jest opublikowana w listopadzie 2020 roku wersja 13 InterCorpu[13], jednak badanie prezentowane w tym artykule bazowało na materiale z wersji 12, opublikowanej w grudniu 2019 roku. Całkowita objętość dostępnej części wersji 12 korpusu to w przybliżeniu 311 milionów słów w wyrównanych obcojęzycznych tekstach w trzonie i 1223 milionów słów w wyrównanych obcojęzycznych tekstach w kolekcjach (Rosen i in., 2019).

Zazwyczaj raz w roku publikowana jest nowa wersja InterCorpu. Z każdą nową wersją rośnie objętość tekstu, mogą pojawiać się nowe języki, zmiana anotacji

[10] Obszernie o różnych korpusach równoległych: Gruszczyńska & Leńko-Szymańska, 2016.

[11] O wykorzystywaniu korpusów równoległych między innymi również w: Cosma i in., 2016; Čermák & Nádvorníková, 2015; Čermáková i in., 2016; Ebeling & Ebeling, 2013; Hebal-Jezierska i in., 2016; Johansson, 2007; Kaczmarska & Rosen, 2013, 2014b; Lewandowska-Tomaszczyk, 2005. Warto zauważyć, iż możliwość wykorzystania przekładów literackich w pracach nad dwujęzycznym słownikiem walencyjnym została zauważona dużo wcześniej (Greń & Rytel-Kuc, 1991).

[12] Termin „wiązanie tekstów" (ang. *alignment*) za: Lewandowska-Tomaszczyk, 2005.

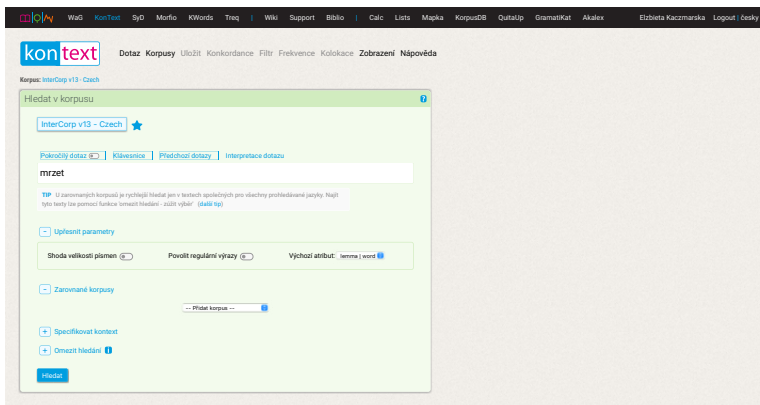
[13] Wcześniejsze wersje (począwszy od wersji 6) są jednak ciągle dostępne.

tekstów czy korekty dotyczące kolekcji[14]. W perspektywie InterCorp ma zawierać również teksty, które nie mają czeskiej wersji; jeszcze jednak nie wiadomo, kiedy to nastąpi. InterCorp straciłby wówczas swoją wyraźną orientację na język czeski (Kaczmarska, 2019; Kaczmarska & Rosen, 2014b).

Dostęp do tekstów

Obecnie InterCorp można przeszukiwać, używając dostępnego od stycznia 2014 roku interfejsu KonText; od kwietnia 2015 roku jest to jedyna wyszukiwarka Czeskiego Korpusu Narodowego (Hebal-Jezierska, 2014; Kaczmarska & Rosen, 2014b). Wcześniej wykorzystywane były również inne interfejsy: Park i NoSketch Engine (Kaczmarska & Rosen, 2014b).

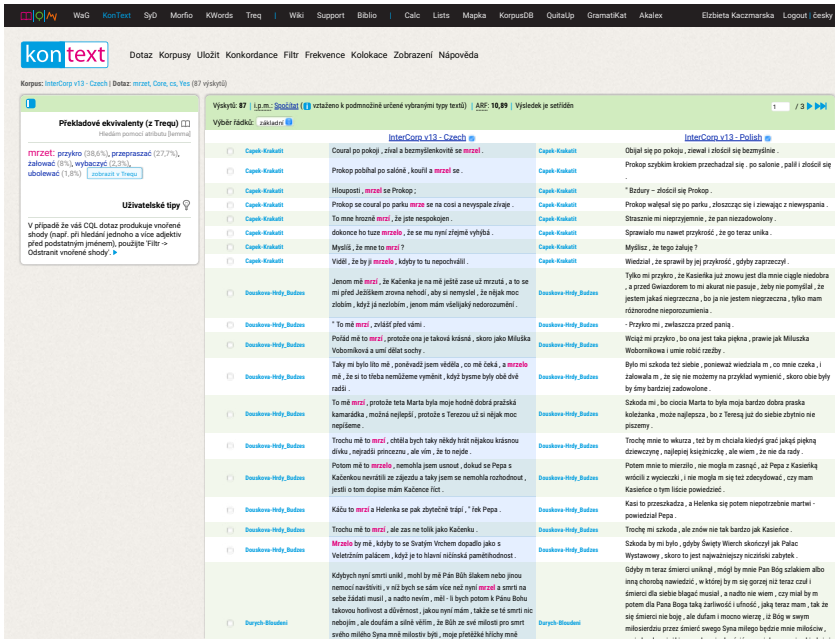
Ikonka KonTextu widoczna jest po otworzeniu strony głównej korpusu (www.korpus.cz)[15]. Wykorzystanie wszystkich funkcji tej wyszukiwarki (i tym samym dostęp do tekstów) możliwe jest po darmowej rejestracji (https://www.korpus.cz/signup). Rejestrując się, użytkownik zyskuje dane dostępowe do wszystkich korpusów ČNK; KonText obsługuje bowiem zarówno korpusy jednojęzyczne, jak i równoległe (Hebal-Jezierska, 2014; Kaczmarska & Rosen, 2014b).



Ryc. 1 Interfejs wyszukiwarki KonText

[14] Niektóre teksty z korpusów Acquis Communautaire i Europarl były częściowo poprawione i wyselekcjonowane, dlatego mogą się różnić od oryginału formą i objętością. Zredukowana została także baza napisów dialogowych (Kaczmarska & Rosen, 2014b).

[15] KonText dostępny jest również bezpośrednio ze strony kontext.korpus.cz



Ryc. 2. Interfejs wyszukiwarki KonText po wyszukaniu paralelnych konkordancji

Polsko-czeska i czesko-polska część InterCorpu

Materiał językowy wykorzystany w prezentowanym badaniu pochodzi przede wszystkim z czesko-polskiej i polsko-czeskiej InterCorpu (Bańczyk i in., 2019), z trzonu korpusu (*core*). Interfejs KonText umożliwia dodatkowe zawężenie wyszukiwania (kluczowe podczas badań nad przekładem) i tak podczas poszukiwań haseł polskich przeszukiwane mogą być wyłącznie teksty napisane w języku polskim i wiązane z tekstami w języku czeskim. Wytyczona w ten sposób część polsko-czeska zawiera 3 461 738 pozycji (w ramach trzonu korpusu, wersja 12 InterCorpu). Analogicznie poświadczenia czeskie wyszukiwane mogą być tylko w tekstach napisanych oryginalnie w języku czeskim i wiązanych z tekstami w języku

polskim[16]. Czesko-polska część zawiera 2 937 087 pozycji. Takie rozwiązanie nakłada jednak na użytkownika konieczność ustalania zawężeń za każdym razem podczas wyszukiwania poszczególnych haseł. Inną możliwością jest tworzenie subkorpusów (Kaczmarska & Rosen, 2014b).

Treq

Treq jest usługą automatycznego generowania słowników przekładowych na podstawie danych zebranych w korpusie równoległym InterCorp, dostępną ze strony Czeskiego Korpusu Narodowego – http://treq.korpus.cz (Rosen & Vavřín, 2015; Škrabal & Vavřín, 2017).

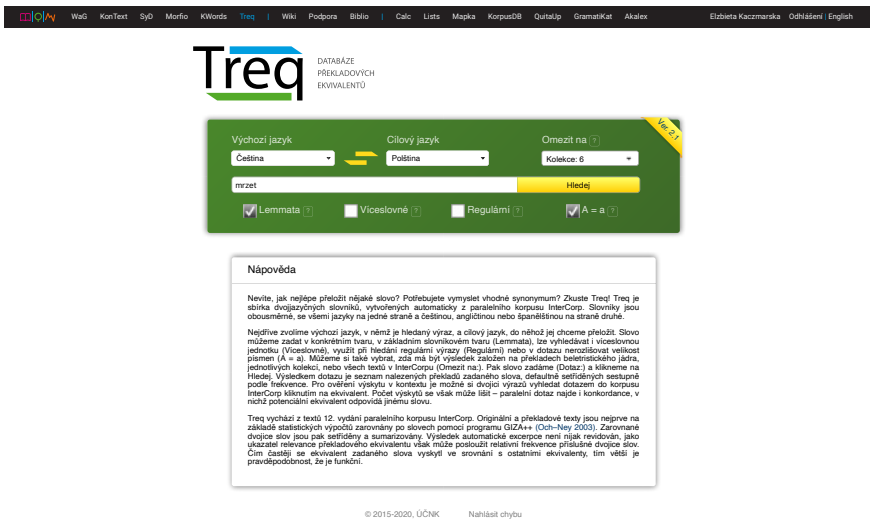
Pomysł automatycznego generowania listy ekwiwalentów zrodził się podczas ustalania polskiego ekwiwalentu czeskiego czasownika *zdát se* w latach 2012–2013 (Kaczmarska, 2012a, 2012b). Badanie ukazało, iż InterCorp oferuje kilkadziesiąt różnych odpowiedników tego czasownika; stwierdzenie tego było jednak możliwe dopiero po wyeksportowaniu wyników do dokumentu Excel, następnie manualnej analizie i ręcznym otagowaniu wszystkich poświadczeń, wreszcie posortowaniu, przefiltrowaniu i przeliczeniu wyników. Było to niezwykle praco- i czasochłonne, niemożliwym więc wydawało się powtórzenie całego procesu dla wszystkich interesujących autorkę czasowników. Jedynym rozwiązaniem, które się nasuwało, było zautomatyzowanie tego procesu. Autorka tego opracowania, wspólnie z dr. inż. Alexandrem Rosenem, podjęła próbę automatycznej ekstrakcji par ekwiwalentów z czesko-polskiego korpusu równoległego, wykorzystując narzędzie GIZA++[17]. Lista par utworzyła swoisty słownik wygenerowany metodą automatyczną[18]. Nie była

[16] Objętość przeszukiwanych tekstów możemy zawęzić poprzez doprecyzowanie, które metadane te teksty mają charakteryzować. Dzieje się to w momencie wpisywania zapytania (Kaczmarska & Rosen, 2014b).

[17] GIZA++ jest programem wolno dostępnym, wytworzonym w celu statystycznego przekładu maszynowego, który zawiera moduł wiązania segmentów na poziomie wyrazu (*word-to-word-alignment*) – http://www.statmt.org/moses/giza/GIZA^{++}.html (Kaczmarska & Rosen, 2013; Och & Ney, 2003).

[18] W trakcie ekstrakcji wykorzystane zostały teksty z korpusu InterCorp (wersja 6). Uwzględniona została jedynie beletrystyka (bez rozróżnienia na czeskie, polskie czy obce oryginały) w tym teksty w języku czeskim: 11 885 milionów słów i teksty w języku polskim 11 860 milionów słów. Wiązanie segmentów (zdań) zostało ograniczone do 1:1; oznacza to, że były wykorzystane tylko segmenty wiązane 1:1, co zwiększyło wiarygodność wiązania na poziomie zdania i wyrazu. Metodą tą wyekstraktowano 8 651 milionów par lematów. Po złączeniu identycznych par (z zachowaniem informacji o ich liczbie) powstało 528 tysięcy haseł dwujęzycznych (Kaczmarska & Rosen, 2013). Podobną metodę wykorzy-

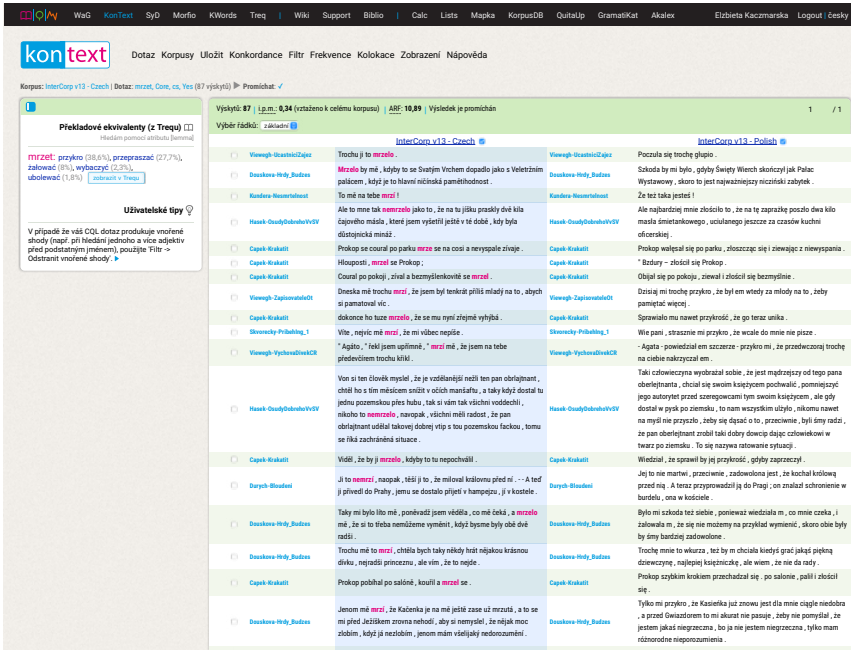
to jednak metoda idealna, gdyż wymagała zaawansowanych znajomości procedur informatycznych i ponownie zajmowała wiele czasu. Poza tym aktualizacja tegoż słownika wymagałaby powtórzenia całej procedury od początku (na nowych danych). Za idealne rozwiązanie autorka przyjęła możliwość generowania takiej listy wprost z korpusu. Pierwszy raz autorka wspomniała o tym podczas wykładu gościnnego w Czeskim Korpusie Narodowym 29 kwietnia 2014 roku. Pomysł zakładał, iż w trakcie wyszukiwania w InterCorpie dwujęzycznych konkordancji na ekranie z wynikami pojawi się okienko – tabelka ze wszystkimi dostępnymi w korpusie ekwiwalentami wraz z danymi o ich liczbie i udziale procentowym. W dyskusji wziął udział dr Michal Křen, który zaproponował utworzenie specjalnej aplikacji, która wykorzystywałaby wiązanie *word-to-word*. Dopiero wyniki wyszukiwań poprzez tę aplikację odsyłałyby do poświadczeń korpusowych. Tak narodził się pomysł stworzenia aplikacji nazwanej później Treq (Rosen & Vavřín, 2015; Škrabal & Vavřín, 2017).



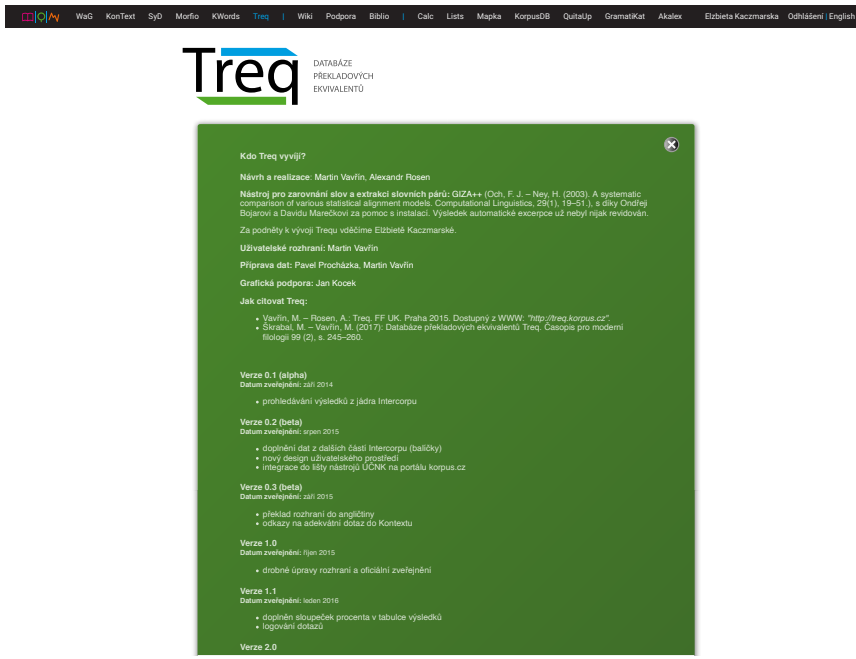
Ryc. 3. Treq – interfejs

stywał Karel Jirásek w swej pracy: *Využití paralelního korpusu InterCorp k získávání ekvivalentů pro chorvatsko-český slovník* (Jirásek, 2011).

Okienko z proponowanymi ekwiwalentami z Treq, pojawiające się po wyświetleniu wyników wyszukiwania w korpusie InterCorp poprzez KonText, zaczęło się pojawiać od wersji 10 InterCorpu; na ilustracji zaznaczone czerwonym kółkiem:

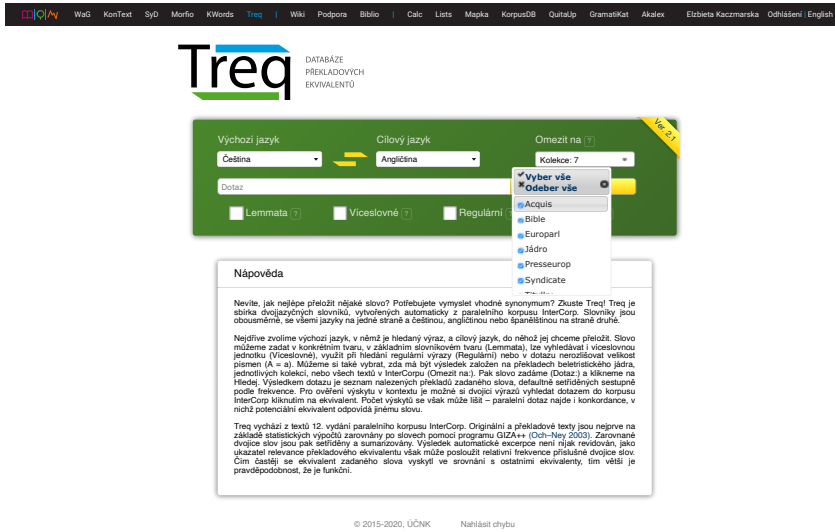


Ryc. 4. Okienko z ekwiwalentami z Treq pojawiające się w interfejsie KonText podczas wyświetlenia wyników wyszukiwania

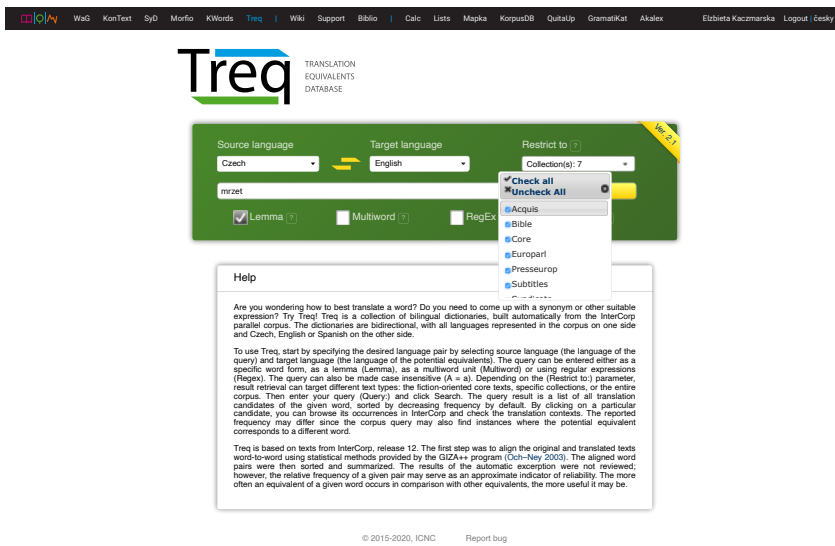


Ryc. 5. Treq – informacja bibliograficzna na stronie internetowej interfejsu

Treq jest narzędziem umożliwiającym generowanie list ekwiwalentów, w których jednym z języków musi być czeski, angielski albo hiszpański, a drugim – którykolwiemk dostępny w tekstach InterCorpu. Oznacza to, że możemy wygenerować np. listę serbskich ekwiwalentów czeskiego słowa i listę czeskich ekwiwalentów serbskiego słowa, ale język polski (podobnie jak wszystkie inne poza czeskim, angielskim i hiszpańskim) wystąpi w kombinacji wyłącznie z językiem angielskim, hiszpańskim i czeskim.



Ryc. 6. Treq – moduł wyszukiwania ekwiwalentów w języku czeskim



Ryc. 7. Treq – moduł wyszukiwania w ekwiwalentów w języku angielskim

Korzystając z tego serwisu, warto jednak wziąć pod uwagę, iż narzędzie to stosuje metodę automatyczną, a wśród odpowiedników mogą znaleźć się sporadycznie również antonimy czy przypadkowe wyrazy. Nie jest więc to typowy słownik, zawierający wyłącznie trafione ekwiwalenty (Kaczmarska, 2019; Kaczmarska & Rosen, 2013). Wykorzystując aplikację Treq, trzeba mieć również świadomość, iż poświadczenia w tym serwisie zliczane są z całego korpusu bez względu na język oryginału. Dużą zaletą jest natomiast automatyczna aktualizacja danych; Treq pobiera listę ekwiwalentów z najnowszej dostępnej wersji InterCorpu.

Próba znalezienia polskiego ekwiwalentu dla czeskiego czasownika *mrzet*

Czasownik *mrzet* w najpopularniejszym polskim słowniku czesko-polskim (Siatkowski & Basaj, 2002) ma kilka odpowiedników: *gniewać*, *złościć*, *mierzić*, *martwić*, *być przykro*, *nie podobać się*, *żałować*, *nudzić*. Klaster ekwiwalentów jest (jak na tradycyjny słownik) dość długi i łączy różne znaczenia. Większość znajduje swoje odniesienia w słowniku języka czeskiego (Havránek i in., 1989), który definiuje *mrzeti* jako: *nemile se dotýkat*, *vyvolávat něčí mrzutou náladu*, *litovat*, *nemít chuť*, *nemít náladu*, *nelíbit se*.

Ekwiwalenty ze słownika dwujęzycznego reprezentują szeroki wachlarz znaczeń; można sądzić, że właściwy odpowiednik może być wskazany dzięki kontekstowi, dlatego ekscerpowane są poświadczenia z InterCorpu. W korpusie równoległym znaleziono 79 przykładów czasownika *mrzet*[19] wraz z wiązanymi segmentami polskimi, w których pojawiło się 30 różnych odpowiedników wyszukiwanego czasownika; ekwiwalenty należą do kilku pól znaczeniowych, między innymi: przykrość [(*być*) *przykro*, *żal*, *wyrzuty sumienia*, *żałować*], smutek (*smucić się*), złość [(*być*) *zły*, *gniewać się*, *drażnić*, *irytować*)], nuda (*nudzić się*), wstyd (*wstydzić się*), niezadowolenie (*szkoda*, *dąsać się*, *przeszkadzać*, *narazić się na nieprzyjemności*), zmartwienie (*martwić się*). Dwa razy czasownik czeski został w tłumaczeniu pominięty (w tabeli – ominięcie).

Większość z polskich odpowiedników to pojedyncze lub podwójne trafienia (wśród nich również czasownik *mierzić*[20]); pojawiają się w polskich segmentach raz albo dwa razy. Dwa ekwiwalenty mają po trzy poświadczenia (*niezadowolony*, *szkoda*), jeden pojawia się w materiale cztery razy (*żałować*) i jeden – pięć razy [(*być*) *zły*].

---

[19] Przeszukiwane były wyłącznie teksty oryginalnie czeskie z trzonu wersji 12 korpusu InterCorp.

[20] Według *Dobrego słownika* czasownik *mierzić* ma wiele znaczeń – budzić wstręt; działać odpychająco; napawać odrazą; odstraszać; odstręczać; wywoływać obrzydzenie; wzbudzać niechęć; zniechęcać; zrażać. Dostęp: https://dobryslownik.pl/slowo/mierzi%C4%87/28273/

Tabela 1. Polskie ekwiwalenty czeskiego czasownika *mrzet* na podstawie czesko-polskiej części korpusu równoległego InterCorp

mrzet	**79**
przykro	30
być zły	5
żałować	4
niezadowolony	3
szkoda	3
drażnić	2
gniewać się	2
martwić	2
przeszkadzać	2
złościć	2
bieda	1
czuć się głupio	1
czuć się gorzej	1
dąsać się	1
irytować	1
mierzić	1
narazić się na nieprzyjemności	1
nie być miło	1
nie podobać się	1
nudzić się	1
wkurzać	1
wstydzić się	1
wyrzuty sumienia	1
zasmucać	1
zdenerwować	1
zmartwić się	1
zmartwiony	1
znudzić się	1
źle	1
żal	1
ominięcie	2

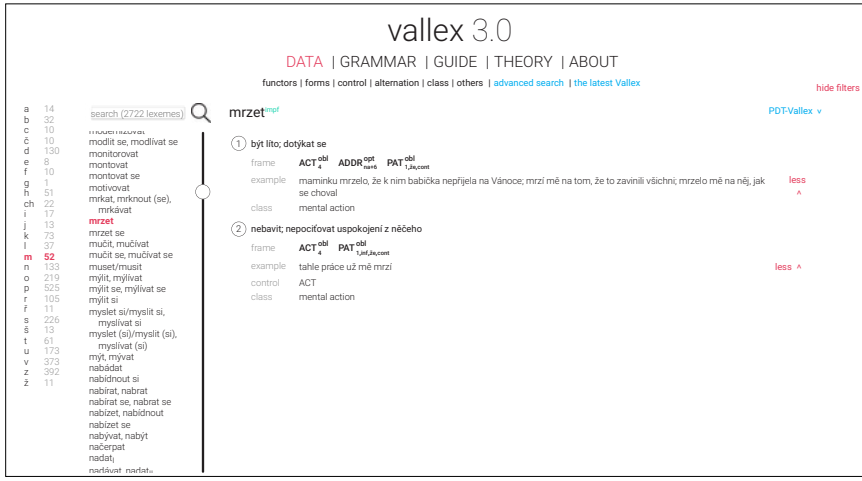
Przeważająca liczba polskich ekwiwalentów łączy się z pojęciem przykrości (30 poświadczeń) – *(być) przykro.*

cs. *Mrzí mě, co jsem řekl, když jsem viděl tvou nedůvěru ve chvíli své největší radosti.*
pl. *Tak mi przykro za to, co powiedziałem, kiedym widział twoją nieufność w chwili mej największej radości!*

cs. *Mrzelo ho, že nemůže poskytnout alespoň tuto útěchu, když už jiná nebyla povolená a vlastně ani možná.*
pl. *Było mu przykro, że nie może udzielić choćby takiej pociechy, skoro już inna nie była dozwolona, a właściwie nawet możliwa.*

cs. *Nedopatření s obědem ho mrzelo.*
pl. *Przykro mu było, że nie dogodził mu obiadem.*

Wybór takiego ekwiwalentu wiąże się jednak ze zmianą struktury zdania: czasownik *mrzet* ma inną charakterystykę walencyjną niż fraza *(być) przykro*, co potwierdzają ich hasła w słownikach walencyjnych.



Ryc. 8. Charakterystyka walencyjna czasownika *mrzet* w słowniku walencyjnym Vallex[21]

Ryc. 9. Charakterystyka walencyjna czasownika *mrzet* w słowniku PDT-Vallex[22]

[21] Dostęp – http://ufal.mff.cuni.cz/vallex/3.5/#/lexeme/mrzet1/0
[22] Dostęp – http://lindat.mff.cuni.cz/services/PDT-Vallex/?verb=mrzet

Czeski czasownik *mrzet* łączy się z frazą nominalną w bierniku [*maminku* (NP_{ACC}) *mrzelo*] oraz frazą zdaniową i bezokolicznikową (*maminku mrzelo, že jste nejedli*; *Alenku mrzelo, jak se choval*; *mrzí ho tam jít*) lub frazą nominalną w mianowniku [*tahle práce* (NP_{NOM}) *mě mrzí*]. Jak ukazuje słownik walencyjny *Walenty*[23], polski odpowiednik *(być) przykro* łączy się z frazą nominalną w celowniku [*było mi* (NP_{DAT}) *przykro*] oraz drugim elementem, który może być realizowany jako fraza zdaniowa i bezokolicznikowa (*przykro mi, gdy zachowujesz się w ten sposób*; *byłoby mi przykro, jeśli to prawda*; *było mu przykro patrzeć na to*) lub fraza nominalna z różnymi komponentami, mi.in. *przez wzgląd na* +NP_{ACC}, *w związku z* + NP_{INS}, *z uwagi na* + NPAcc (*jest mi przykro przez wzgląd na kibiców*; *bardzo nam przykro w związku z zaistniałą sytuacją*; *jest mi bardzo przykro z uwagi na śmierć Tomka*)[24]. Taka różnica ram walencyjnych może utrudnić przekład.

Jeszcze szerszy wachlarz ekwiwalentów oferuje Treq. Automatycznie wygenerowana lista odpowiedników kontekstowych obejmuje około 200 różnych pozycji; niektóre są jednak błędnie wiązane (np. *tak*, *naprawdę*, *chcieć*).

Tabela 2. Polskie ekwiwalenty czeskiego czasownika *mrzet* na podstawie narzędzia Treq[25]

mrzet	**3050**
przykro	1174
przepraszać	846
żałować	239
wybaczyć	74
ubolewać	67
szkoda	49
żal	40
martwić	30
pożałować	27
tak	20
przykrość	20
słyszeć	18
przeprosić	17

[23] Dostęp – http://walenty.ipipan.waw.pl

[24] Frazy te oddają źródło owej „przykrości". Podobnie możemy interpretować konstrukcje typu: *Jest nam przykro po tej porażce* – chodzi tu pierwotnie o ustalenie przyczyny „bycia przykro" a nie czasu. Przykład ten jest również przytaczany w słowniku Walenty.

[25] W tabelce tej zostały przedstawione tylko ekwiwalenty, które pojawiły się przynajmniej 10 razy w poświadczeniach.

mrzet	**3050**
zmartwić	16
naprawdę	13
chcieć	13
zły	12
rozczarować	11
źle	11
współczuć	11
przykry	10

Lista ekwiwalentów wygenerowana dzięki narzędziu Treq obejmuje czasowniki należące do różnych pól znaczeniowych podobnie jak w przypadku zestawu pochodzącego z InterCorpu; jest to częściowo przewidywalne, ponieważ Treq czerpie z zasobów InterCorpu, wyszukiwania nie można jednak w tym przypadku ograniczyć do czesko-polskiej części korpusu. Na liście tej (tabela 2) najwyższą pozycję zajmuje *przykro* – podobnie jak na liście konstruowanej na podstawie czesko-polskiej części korpusu równoległego InterCorp (tabela 1). Na drugiej pozycji znajduje się czasownik *przepraszać* (niżej w tabeli także *przeprosić* – 17 poświadczeń); ekwiwalenty te nie pojawiły się w ogóle w tabeli 1.

cs. *To mě moc mrzí, slečno.*
pl. *Bardzo przepraszam, panienko.*

cs. *Mrzí mě, že jsem propásl tvůj proces.*
pl. *Przepraszam, że nie przyszedłem na proces.*

cs. *A mrzí mě, že jsem se k tobě zachoval tak ohavně.*
pl. *Przepraszam, że zachowywałem się w stosunku do ciebie obrzydliwie.*

Na trzeciej pozycji znalazł się czasownik *żałować*; to samo miejsce zajmuje ten czasownik również w tab. 1.

cs. *Mrzí mě, že fotografie nedokázala zachytit její podobu v barvách.*
pl. *Żałuję, że fotografia nie może pokazać kolorytu jej urody.*

cs. *Opravdu nás mrzí, že odjíždíte, madam.*
pl. *Szczerze żałujemy, że pani wyjeżdża.*

cs. *Bude ho to pěkně mrzet.*
pl. *Jeszcze będzie żałował, że mi nie pomógł.*

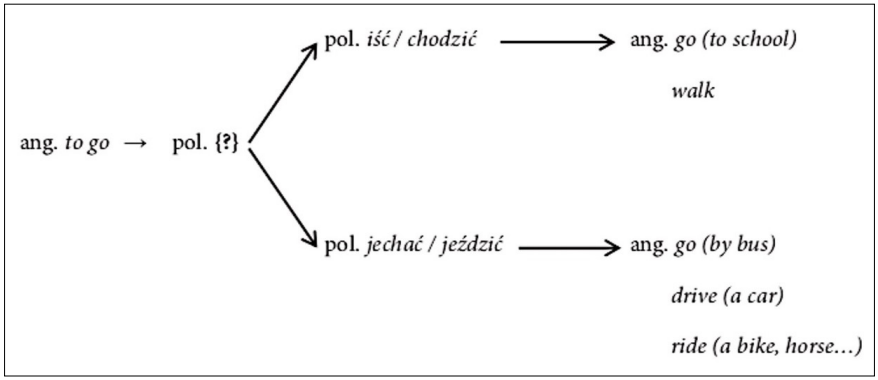
Obie listy ekwiwalentów ukazują, iż najczęstszymi ekwiwalentami są (*być*) *przykro*, *żałować* oraz *przepraszać*. Oba zestawy zawierają również inne odpowiedniki, które niosą odmienne znaczenia, a także mogą modyfikować strukturę przekładu w stosunku do oryginału. Słownik języka czeskiego prezentował jedynie dwa znaczenia; przy przekładzie na język polski czasownika *mrzet* tłumacze wykorzystują jednak szereg różnych jednostek leksykalnych, należących do różnych pól znaczeniowych. Czasownik *mrzet* jawi się jako jeden z wieloznacznych czasowników, dla których ustalenie polskiego ekwiwalentu jest kłopotliwe. Pierwszym problemem może być już samo określenie, co w danym kontekście badany czasownik znaczy; kwestia wskazania polskiego ekwiwalentu jest sprawą wtórną. Warto w tym miejscu odnieść się do słów Elżbiety Tabakowskiej (Tabakowska, 1991, 1995), która postrzega przekład jako grę konceptualizacji i obrazowania.

> Kopia nigdy nie będzie wierna do końca – po pierwsze dlatego, że ten sam subiektywizm w tworzeniu obrazów świata, który pozwala na niezliczone warianty obrazowania, wyklucza jednocześnie identyczność interpretacji. Po drugie zaś dlatego, że techniki dostępne twórcy oryginału (językowe konwencje) mogą nie mieć odnośników w języku kopisty (Tabakowska, 1991, s. 60).

W przypadku badanego czasownika odtwarzamy mapę emocji i próbę ich oddania w języku polskim. Współgra to teorią rekonceptualizacji (Lewandowska-Tomaszczyk, 2010a, 2010b), która zakłada, iż tłumacząc wypowiedzi, dokonuje się ponownej konstrukcji danych pojęć. Dzieje się tak w przykładzie przytaczanym przez Barbarę Lewandowską-Tomaszczyk (Lewandowska-Tomaszczyk, 2010b): *I'd like Tom to come earlier*. Język polski nie umożliwia pojawiania się w tłumaczeniu bezokolicznika; wymaga obecności zdania podrzędnego, a w konsekwencji i wskaźników czasu gramatycznego, aspektu, liczby i rodzaju. Oprócz tego mówiący w języku polskim musi „ustalić", czy Tom idzie, czy jedzie. W rezultacie można wskazać szereg potencjalnych ekwiwalentnych zdań; każde z nich dotknięte jednak będzie rekonceptualizacją:

> Chciałabym / chciałbym, żeby Tomek przyjechał / przyszedł wcześniej.

Istota rekonceptualizacji związana jest z pojęciem asymetrii semantycznej, której przykład prezentuje Barbara Lewandowska-Tomaszczyk:



Ryc. 10. Przykład asymetrii semantycznej – ang. *displacement of senses*

Cykle rekonceptualizacji obserwuje się również w procesie przekładu pomiędzy dwoma bliskimi językami; dotyczy to także języka czeskiego i polskiego (Kaczmarska, 2019). Wskazać tu można choćby rekonceptualizację „zubożeniową", jak np. w przypadku analizowanego czeskiego czasownika *mrzet*.

Ryc. 11. Przykład asymetrii semantycznej czasownika *mrzet* i jego odpowiedników w języku polskim

Wieloznaczny (dla polskiego odbiorcy) czasownik czeski podczas przekładu na język polski musi zostać ujednoznaczniony, a język polski dysponuje całym zbiorem możliwych odpowiedników czasownika *mrzet*. Struktury ze słowem *przykro* (najczęściej z zaimkiem *mi*) czy czasownik *żałować*, zarówno sugerowane przez słownik, jak i obecne wśród ekwiwalentów korpusowych, są jednak bliższe znaczeniowo czasownikowi *litovat* czy frazie *být líto*. Wobec czasownika *mrzet* są ekwiwalentne tylko w części przypadków. Czasownik *mrzet* oprócz żalu zawiera także nutę złości; z tego powodu bywa tłumaczony na polski również jako *gniewać*, *drażnić* czy *irytować*. Pomiędzy czasownikami *żałować* i *drażnić* jest jednak w języku polskim semantyczna przepaść.

Ostateczną decyzję o kształcie tłumaczenia i doborze ekwiwalentów podejmuje tłumacz, jednak skorzystanie z korpusu równoległego czy narzędzia Treq może to zadanie w znaczący sposób ułatwić.

BIBLIOGRAFIA

Bańczyk, Ł., Dybalska, R., & Vavřín, M. (2019). *Korpus InterCorp – polština, verze 12 z 12. 12. 2019*. Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy.

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427. https://doi.org/10.1075/ijcl.17.3.0 5cer

Čermák, P., & Nádvorníková, O. (2015). *Románské jazyky a čeština ve světle paralelních korpusů*. Karolinum.

Čermáková, A., Chlumská, L., & Malá, M. (Red.). (2016). *Jazykové paralely*. Nakladatelství Lidové noviny.

Cosma, R., Cristea, D., Kupietz, M., Tufiş, D., & Witt, A. (2016). DRuKoLA – towards contrastive German-Romanian research based on comparable corpora. W N. Calzolari, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Red.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia* (ss. 28–32). European Language Resources Association (ELRA).

Dziob, A., & Piasecki, M. (2018). Implementation of the verb model in plWordNet 4.0. W F. Bond, T. Kuribayashi, C. Fellbaum, & P. Vossen (Red.), *Proceedings of the 9th Global Wordnet Conference* (ss. 114–123). The Global WordNet Association.

Ebeling, J., & Ebeling, S. O. (2013). *Patterns in contrast*. John Benjamins. https://doi.org/10.1075/scl.58

Fellbaum, C. (Red.). (1998). *WordNet: An electronic lexical database*. MIT Press. https://doi.org/10.7551/mitpress/7287.001.0001

Greń, Z., & Rytel-Kuc, D. (1991). Wykorzystanie przekładów literackich w pracy nad dwujęzycznym słownikiem walencyjnym. W Z. Rudnik-Karwatowa, H. Běličova, & G. Nieszczimienko (Red.), *Problemy teoretyczno-metodologiczne badań konfrontatywnych języków słowiańskich* (ss. 69–78). Instytut Słowianoznawstwa Polskiej Akademii Nauk.

Gruszczyńska, E., & Leńko-Szymańska, A. (Red.). (2016). *Polskojęzyczne korpusy równoległe / Polish-language parallel corpora*. Instytut Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Havránek, B., Bělič, J., Helcl, M., Jedlička, A., Křístek, V., & Trávníček, F. (Red.). (1989). *Slovník spisovného jazyka českého* (2. wyd., T. 3). Academia.

Hebal-Jezierska, M. (2014). Praktyczny przewodnik po korpusie języka czeskiego. W M. Hebal-Jezierska (Red.), *Praktyczny przewodnik po korpusach języków słowiańskich* (ss. 29–54). Wydział Polonistyki Uniwersytetu Warszawskiego.

Hebal-Jezierska, M., Kaczmarska, E., & Rosen, A. (2016). 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. W E. Gruszczyńska & A. Leńko-Szymańska (Red.), *Polskojęzyczne korpusy równoległe / Polish-language parallel corpora* (ss. 41–65). Instytut Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Janz, A., Kędzia, P., & Kaszewski, D. (2018). *Word Sense Disambiguation tool WoSeDon*. http://hdl.handle.net/11321/540

Jirásek, K. (2011). Využití paralelního korpusu InterCorp k získávání ekvivalentů pro chorvatsko-český slovník. W Čermák, F. (Red.), *Korpusová lingvistika Praha 2011: 1 – InterCorp* (ss. 45–55). Nakladatelství Lidové noviny, Praha.

Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.26

Kaczmarska, E. (2012a). Czeski czasownik *zdát se* w przekładzie na język polski (na podstawie badań z wykorzystaniem czesko-polskiego korpusu równoległego InterCorp). *Studia z Filologii Polskiej i Słowiańskiej*, *47*, 247–261. https://doi.org/10.11649/sfps.2012.012

Kaczmarska, E. (2012b). Searching for equivalents on the basis of Czech-Polish parallel corpus: The case of the verb *zdát se*. W P. Karag'ozov, K. Bakhneva, V. Geshev, I. Khristova, & M. Mladenova (Red.), *Vreme i istoriia v slavianskite ezitsi, literaturi i kulturi: Ezikoznanie* (ss. 238–245). Universitetsko Izdatelstvo Sv. Kliment Okhridski.

Kaczmarska, E. (2016). O dwóch jednostkach leksykalnych będących wykładnikami negatywnych stanów emocjonalnych i ich polskich ekwiwalentach: Analiza na materiale z korpusu równoległego InterCorp. W E. Gruszczyńska & A. Leńko-Szymańska (Red.), *Polskojęzyczne korpusy równoległe / Polish-language parallel corpora* (ss. 227–248). Wydział Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Kaczmarska, E. (2019). *Metody ustalania ekwiwalentów czasowników wyrażających stany emocjonalne w przekładzie czesko-polskim na materiale z korpusu równoległego InterCorp*. Wydział Polonistyki Uniwersytetu Warszawskiego.

Kaczmarska, E., & Rosen, A. (2013). Między znaczeniem leksykalnym a walencją: Próba opracowania metody ekstrakcji ekwiwalentów na podstawie korpusu równoległego. *Studia z Filologii Polskiej i Słowiańskiej*, *48*, 103–121. https://doi.org/10.11649/sfps.2013.007

Kaczmarska, E., & Rosen, A. (2014a). Czego nie można wyrazić w języku polskim, czyli o leksykalnych w nim brakach. *Polonica*, *34*, 53–66.

Kaczmarska, E., & Rosen, A. (2014b). Praktyczny przewodnik po korpusie równoległym In- terCorp. W M. Hebal-Jezierska (Red.), *Praktyczny przewodnik po korpusach języków słowiańskich* (ss. 207–231). Wydział Polonistyki Uniwersytetu Warszawskiego.

Kędzia, P., Piasecki, M., & Orlińska, M. (2016). *WoSeDon: CLARIN-PL digital repository*. Wrocław University of Technology. https://clarin-pl.eu/dspace/handle/11321/290

Lewandowska-Tomaszczyk, B. (1984). *Conceptual analysis, linguistic meaning, and verbal interaction*. Wydawnictwo Uniwersytetu Łódzkiego.

Lewandowska-Tomaszczyk, B. (2010a). Nowe spojrzenie na przekład: Podobieństwo, granice ekwiwalencji i rekonceptualizacja, *Lingwistyka Stosowana*, *3*, 9–31.

Lewandowska-Tomaszczyk, B. (2010b). Re-conceptualization and the emergence of discourse meaning as a theory of translation. W B. Lewandowska-Tomaszczyk & M. Thelen (Red.), *Meaning in translation* (ss. 105–147). Peter Lang.

Lewandowska-Tomaszczyk, B. (2013). *Komunikacja i konstruowanie znaczeń w przekładzie: Referat wygłoszony na konferencji „Zbliżenia: Językoznawstwo"* [Niepublikowany referat]. Konin.

Lewandowska-Tomaszczyk, B. (Red.). (2005). *Podstawy językoznawstwa korpusowego*. Wydawnictwo Uniwersytetu Łódzkiego.

Łaziński, M. (2014). Praktyczny przewodnik po korpusach równoległych: Wiadomości wstępne: Korpus ParaSol i Korpus polsko-rosyjski Uniwersytetu Warszawskiego. W M. Hebal-Jezierska (Red.), *Praktyczny przewodnik po korpusach języków słowiańskich* (ss. 199–206). Wydział Polonistyki Uniwersytetu Warszawskiego.

Łaziński, M., Kuratczyk, M., Orekhov, B., & Słobodjan, E. (2012). The Polish-Russian Parallel Corpus and its application in the linguistic analysis. *Prace Filologiczne*, *63*, 209–218.

Maziarz, M., Piasecki, M., & Rudnicka, E. (2014). Słowosieć – polski wordnet: Proces tworzenia tezaurusa. *Polonica*, *34*, 79–97.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51. https://doi.org/10.1162/089120103321337421

Piasecki, M., Kędzia, P., & Orlińska, M. (2016). plWordNet in word sense disambiguation task. W V. B. Mititelu, C. Forăscu, C. Fellbaum, & P. Vossen (Red.), *Proceedings of the 8th Global Wordnet Conference* (ss. 280–289). The Global WordNet Association.

Rosen, A. (2016). InterCorp: A look behind the façade of a parallel corpus. W E. Gruszczyńska & A. Leńko-Szymańska (Red.), *Polskojęzyczne korpusy równoległe: Polish-*

language parallel corpora (ss. 21–40). Wydział Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Rosen, A., & Vavřín, M. (2014). *Korpus InterCorp: Čeština, verze 7 z 19.12.2014*. Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy. http://www.korpus.cz

Rosen, A., & Vavřín, M. (2015). *Treq*. Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy. http://treq.korpus.cz

Rosen, A., Vavřín, M., & Zasina, A. J. (2019). *Korpus InterCorp, verze 12 z 12.12.2019*. Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy. http://www.korpus.cz

Siatkowski, J., & Basaj, M. (2002). *Słownik czesko-polski*. Wiedza Powszechna.

Škrabal, M., & Vavřín, M. (2017). Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii*, *99*(2), 245–260.

Tabakowska, E. (1991). Przekład i obrazowanie. *Arka*, *34*, 52–61.

Tabakowska, E. (1995). *Gramatyka i obrazowanie: Wprowadzenie do językoznawstwa kognitywnego*. Polska Akademia Nauk.

von Waldenfels, R. (2012). ParaSol: Introduction to a Slavic parallel corpus. *Prace Filologiczne*, *63*, 293–301.

InterCorp i Treq jako narzędzia ułatwiające wyszukiwanie ekwiwalentów

Abstrakt

Artykuł poświęcony jest narzędziom ujednoznaczniającym, które mogą być pomocne w rozwiązywaniu problemu rozumienia i tłumaczenia wieloznacznych słów, w szczególności czasowników oznaczających stany emocjonalne i psychiczne. Artykuł przedstawia InterCorp, wielojęzyczny, równoległy korpus, który obecnie obejmuje 41 języków; zawiera też porady techniczne dotyczące dostępu do tekstu i wyszukiwania. Przedstawia również ideę i genezę aplikacji Treq oraz mechanizm jej działania. Wykorzystując InterCorp i Treq, autorka przeprowadza analizę mającą na celu znalezienie polskiego odpowiednika czeskiego czasownika *mrzet*.

Słowa kluczowe: korpus równoległy; InterCorp; Treq; ekwiwalent, przekład, język czeski, język polski

InterCorp and Treq as Tools Facilitating the Search for Equivalents

Abstract

This article is devoted to disambiguation tools which may be helpful in solving the problem of understanding and translating ambiguous words, particularly verbs denoting emotional and mental states. The study offers an introduction to InterCorp, a multilingual parallel corpus currently covering 41 languages, and provides technical advice on text access and search. It also outlines the idea and genesis of the application called Treq as well as the mechanism of its operation. Using InterCorp and Treq, the author conducts an analysis aiming to find the Polish equivalent for the Czech verb *mrzet*.

Keywords: parallel corpus; InterCorp; Treq; equivalent; translation; Czech language; Polish language