„Bohemistyka” 2020, nr 4, ISSN 1642–9893

Andrzej CHARCIAREK
Uniwersytet Ślaski w Katowicach

DOI: 10.14746/bo.2020.4.4

Narzędzie Treq w procesie ustalania polsko-czeskich par przekładowych

Keywords: Treq, InterCorp parallel corpus, translation equivalence, translation pairs, Czech, Polish
Słowa kluczowe: Treq, korpus równoległy InterCorp, ekwiwalencja przekładowa, pary przekładowe, jezyk czeski, język polski

Abstract

The article is devoted to the use of the Treq corpus tool in the proces of determining language pairs in Polish-Czech translation. It provides a characteristics of textual resources found in the InterCorp parallel corpus and focuses on their usability in translation of various text genres. Additionally, article describes different functionalities of the TreQ tool. In order to illustrate its advantages and disadvantages the author made an attemt to determine Czech equivalents for the Polish word ciacho.

Niniejszy artykuł poświęcony jest wykorzystaniu korpusowego narzędzia Treq w procesie ustalania polsko-czeskich par przekładowych. Scharakteryzowano zasoby tekstowe polsko-czeskiego korpusu równoległego InterCorp pod względem ich przydatności w tłumaczeniu określonych gatunków tekstów. Omówiono funkcje narzędzia Treq. Na konkretnym przykładzie ustalenia czeskich ekwiwalentów przekładowych dla polskiego leksemu ciacho pokazano zarówno walory, jak i mankamenty aplikacji Treq.

Z bazy ekwiwalentów przekładowych Treq (http://treq.korpus.cz) można korzystać począwszy od 8. wersji korpusu równoległego InterCorp (https://intercorp.korpus.cz), którego kolejne wersje co roku udostępniane są użytkownikom. Pierwotnie aplikacja Treq (w wersji 8. InterCorp) umożliwiała zestawianie par przekładowych dla języka czeskiego, przy czym można było zestawiać tylko jednostki jednowyrazowe (lemmy lub formy). Duże zmiany przyniosła 10. wersja InterCorp, w której za pomocą nowej wersji Treq 2.0 można było już zestawiać pary przekładowe nie tylko z językiem czeskim, ale także z angielskim. Nie to jednak było największą zmianą w pracy z tą aplikacją. Jej nowa wersja pozwalała na zestawianie ze sobą jednostek wielowyrazowych, a w procesie ich wyszukiwania na używanie wyrażeń regularnych.

Baza ta jest powiązana z tekstami wielojęzycznego korpusu równoległego InterCorp. Obecnie program ten wykorzystuje zasoby tekstowe 9. wersji tego korpusu z 2016 roku. Jak z tego wynika, Treq nie korzysta z tekstów zamieszczanych w corocznie udostępnianych, systematycznie zwiększających swoje zasoby, kolejnych wersjach InterCorp[1]. Dla przykładu polska 9. wersja InterCorp liczy prawie 84 mln słów, podczas gdy najnowsza 12. wersja – ponad 87,5 mln. Różnicę między oboma wersjami stanowią również przekłady Pisma Świętego Starego i Nowego Testamentu, które zostały włączone do zasobów korpusowych w 10. oraz następnych wersji InterCorp.

Co oczywiste, optymalnym rozwiązaniem byłoby powiązanie Treq z najnowszą wersją korpusu i wykorzystanie jego największych zasobów tekstowych, niemniej wielkość polsko-czeskiego InterCorp i tak stwarza dobre warunki do pracy translatorskiej i leksykograficznej. Podkreślam to szczególnie, ponieważ zasoby tekstów w języku polskim są jednymi z największych i najbardziej zróżnicowanych w całym korpusie Intercorp[2]. W 12. jego wersji znajdziemy literaturę piękną (ponad 25 mln wyrazów), teksty prawne z korpusu Acquis Communautaire (niespełna 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). Jedyną brakującą kolekcją są teksty publicystyczne ze strony Project Syndicate, które nie są tłumaczone na język polski.

[1] Ostatnia 12. wersja korpusu równoległego InterCorp została udostępniona użytkownikom w 2019 roku.

[2] Większe zasoby od języka polskiego w InterCorp mają języki: angielski, hiszpański, niderlandzki, francuski i portugalski.

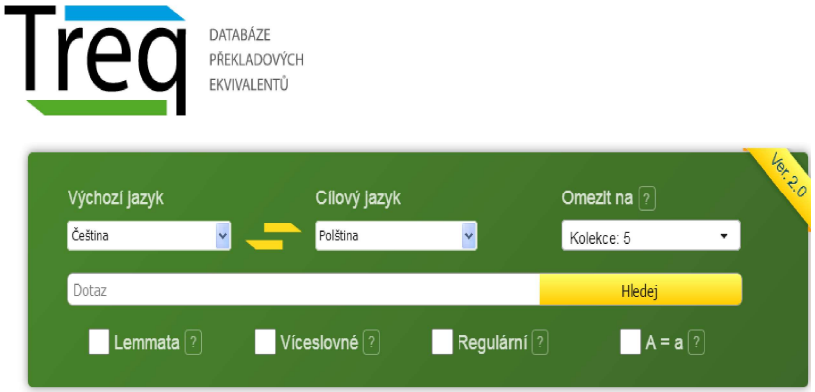
Tak więc, jak już wspomniałem, w przypadku programu Treq dysponujemy nieco mniejszymi zasobami korpusowymi 9. wersji InterCorp.

Treq to w praktyce słowniki przekładowe sporządzone automatycznie na podstawie analizy kwantytatywno-kwalitatywnej tekstów zgromadzonych w korpusie. Teksty oryginalne (w mniejszości) i przekłady (w większości) zostały wyrównane automatycznie na poziomie wyrazów (*word alignment*) za pomocą programu GIZA++ (Och, Ney 2003). Pozyskane w ten sposób pary wyrazów utworzyły listy frekwencyjne uporządkowane według liczby wystąpień w korpusie. Zdaniem autorów Treq, choć pozyskane wyniki nie były poddane żadnej weryfikacji, to ekwiwalenty o najwyższej frekwencji można uznać za trafne. Wygenerowane listy frekwencyjne nie prezentują tylko liczbę poświadczeń ewentualnych ekwiwalentów, ale także przedstawiają ich procentowy udział w stosunku do wszystkich możliwych odpowiedników w tekstach korpusowych – w kolejności od ekwiwalentu z największą frekwencją do ekwiwalentu z najmniejszą.

Użytkownik Treq uzyskuje zatem dostęp do zbioru ekwiwalentów przekładowych wyekscerpowanych z konkretnych tekstów, do których uzyskuje wgląd za pośrednictwem odnośnika hipertekstowego, co umożliwia ich dodatkową weryfikację[3]. W porównaniu ze słownikiem przekładowym czy to w wersji papierowej, czy na nośniku elektronicznym, Aplikacja Treq zmienia diametralnie sposób poszukiwania ekwiwalentów. W przypadku tego pierwszego nierzadko mamy do czynienia z materiałem preparowanym, nieautentycznym, nietrafnymi decyzjami ekwiwalentatyzacyjnymi czy błędami powielanymi z opracowań leksykograficznych poprzedników. Oczywiście, wielu autorom słowników przekładowych nie można odmówić rzetelności badawczej, których dowód stanowią liczne szczegółowe mikroanalizy leksykograficzne zarówno zjawisk językowych częstych, jak i rzadkich. Co do tych ostatnich to ma się wrażenie, że uwaga im poświęcana w słownikach jest niekiedy nadmierna. Aplikacja Treq, jeśli traktować ją jako dokumentacyjny słownik przekładowy, a która w istocie nim jest, stwarza odmienne warunki pozyskiwania odpowiedników przekładowych. Nie jest to podstawowy program do wyszukiwania ekwiwalentów przekładowych w Intercorp, tym pozostaje niezmiennie Kontext. Niemniej Treq ma sporo do zaoferowania odnośnie do tłumaczenia czy opisu leksykograficznego. Oczywiście, pod warunkiem prawidłowego formułowania zapytań i właściwego korzystania z zasobów korpusowych. Pomijam tu kwestię ewentualnych błędów wynikających z zawodności automatycznych funkcji wyszukiwawczych czy pomyłek w anotacji lingwistycznej. Wydaje się, że użytkownicy nie w pełni korzystają z możliwości korpusów, którymi się posługują. Wynika to z dwu przyczyn. Po pierwsze, nieznajomości zasobów korpusowych, a po drugie, ograniczonego wykorzystywania funkcjonalności narzędzi wyszukiwawczych.

Przejdźmy zatem do opisu bazy ekwiwalentów przekładowych Treq i spójrzmy nań jako na narzędzie wspomagające proces przekładu polsko-czeskiego/czesko-polskiego lub służące do opisu leksykograficznego.

Interfejs stanowi powielenie modelu stworzonego dla całego Czeskiego Korpusu Języka Czeskiego, czyli jest w dwu wersjach językowych – czeskiej i angielskiej.



[3] Szczegółowy opis Treq zna j dzie my w artykule M. Škrabala i M Vavřína Databáze překladových ekvivalentů Treq.

Autorzy z pewnością założyli, iż użytkownikami Treq mogą być zarówno specjaliści (tłumacze, leksykografowie), jak i niespecjaliści (uczniowie, studenci) i dlatego aplikację cechuje prostota i przyjazność. Łatwość użycia Treq ma też i wadę – uniemożliwia poszerzoną analizę wieloczynnikową.

Przeszukiwanie w zasobach korpusowych może być przeprowadzone na wybranych kolekcjach (Acquis Communautaire, Europarl, VoxEurop, OpenSubtitles***),*** co pozwala na rezygnację z określonych gatunków tekstów (funkcja *omezit na*). Przy czym możliwa jest także rezygnacja z jądra InterCorp, czyli tekstów beletrystycznych, które tworzą w praktyce piąty zbiór. Zawężenie materiału korpusowego poddanego analizie skutkuje, o czym należy pamiętać, zmniejszeniem liczby poświadczeń, które uniemożliwiają ich wiarygodną interpretację. Teksty zgromadzone w InterCorp nie odzwierciedlają wszystkich wzorców gatunkowych, ale niektóre z nich już tak.

Literatura piękna wnosi do korpusów teksty najbardziej zróżnicowane pod względem gatunkowym i odmianowym, będące pewnym wzorcem normatywnym[4]. W nich znajdziemy także stylizację na mowę potoczną, która nie jest materiałem optymalnym, ale z pewnością w korpusie równoległym przydatnym. Fakt ten odzwierciedla się w pozyskanych poświadczeniach, które mają jeszcze jedną istotną zaletę – w odróżnieniu od pozostałych kolekcji InterCorp wskazywane pary przekładowe są z reguły trafne, a to głównie dzięki temu, że teksty literackie zostały wyrównane ręcznie.

Duże zasoby polskiego InterCorp stanowią amatorskie napisy filmowe z platformy OpenSubtitles[5]. To zasoby istotne dla ustanawiania par przekładowych, zawierające jednostki używane w komunikacji codziennej, głównie nieoficjalnej. Ze względu na ograniczenia czasowo-przestrzenne, które mają wpływ na ich ostateczną formę, wymagają one od użytkownika szczegółowej analizy ze świadomością faktu, że ich postać może nie w pełni odzwierciedlać języka mówionego. Niemniej rekompensują one w znacznej mierze brak tekstów konwersacyjnych wypowiadanych w sytuacji nieoficjalnej, które są deficytowe nie tylko w korpusach równoległych, ale także jednojęzycznych. Wartość napisów jest tym większa, że tworzone są głównie przez osoby młode, biorące aktywny udział w komunikacji w mediach elektronicznych, które w większym stopniu niż literatura piękna kiedyś kształtują ich kompetencje językowe. W nich znajdziemy mowę potoczną i slang, które w coraz większym stopniu oddziałują na komunikację codzienną. Nawiasem mówiąc, należy sobie uświadomić fakt, że korpusy języka mówionego zwykle są częścią narodowych korpusów jednojęzycznych, jako równoległe w zasadzie nie istnieją.

Związek z językiem mówionym wykazują teksty informacyjne i publicystyczne z portalu VoxEurope. To, że tematycznie jako takie na ogół szybko się dezaktualizują, dla wartości zasobów korpusowych nie ma w praktyce znaczenia. Rejestrują one najnowsze zjawiska dokonujące się w języku, które następnie upowszechniają się za pośrednictwem mediów tradycyjnych i elektronicznych. Taka sytuacja ma miejsce choćby z nowymi frazemami, które dzięki wysokiej częstości użycia stopniowo nabierają statusu normy.

Kolekcję Europarl stanowią sprawozdania z obrad Parlamentu Europejskiego, będące zapisem wypowiedzi ustnych parlamentarzystów w sytuacji oficjalnej. To bogaty i zróżnicowany materiał, zawierający nie tylko wypowiedzi informacyjne, ale także komentarze i polemiki polityków. W kolekcji Europarl znajdziemy jednostki językowe przynależne nie tylko do dyskursu debaty publicznej, ale także do dyskursu profesjonalnego i potocznego.

Teksty prawne z korpusu Acquis Communautaire dopełniają polsko-czeski moduł korpusu InterCorp. Zgromadzone w nim tłumaczenia tekstów prawnych, będących gatunkiem tekstów specjalistycznych, dotyczą prawa wspólnotowego. W tym przypadku zasoby korpusowe mogą posłużyć jako źródło ekwiwalentów przekładowych

[4] W środowisku językoznawczym nie ma zgodności co do uznania danych korpusowych za podstawę ocen poprawnościowych.

[5] Opisałem je szczegółowo w artykule *Korpus równoległy InterCorp w leksykografii przekładowej – możliwości i ograniczenia*, „Roczniki Humanistyczne", 2019, s. 79–92.

regulowanych nie normą, ale uzusem (podobnie jak w przypadku umów międzynarodowych).

Jak można zauważyć, zasoby polsko-czeskiego InterCorp, z których korzysta Treq, zawierają tylko niektóre gatunki tekstów użytkowych, nie rejestrują praktycznie wcale gatunków internetowych (np. blogów, forów internetowych, czatów itp.) czy tekstów religijnych. Nic nie ma w tym dziwnego, ponieważ pewne typy tekstów tłumaczone dotąd nie były i najpewniej nie będą, a warunkiem przy tworzeniu korpusu równoległego InterCorp jest istnienie oryginału i przekładu, o czym pisze František Čermák:

> Korpus InterCorp spojuje dohromady přes třicet jazyků, které jsou všechny propojené přes centrální češtinu; lze v nich však hledat, pokud příslušné příklady existují, i tak, že se čeština vynechá a srovnávají se jiné jazyky mezi sebou přímo. To ovšem předpokládá dostupnost originálních i překladových textů a ty nemusejí být dostupné vždy a snadno, anebo nemusejí vůbec existovat. Proto je tvorba InterCorpu v zásadě pragmatické povahy (nemůže v něm být to, co nebylo přeloženo) (*Čermák* 2017, s. 82).

Można sądzić, że w przyszłości w zasobach InterCorp znajdą się także inne, dotychczas nienotowane teksty użytkowe, np. instrukcje obsługi urządzeń, opisy sprzętu, dokumentacje maszyn itp. Ma to związek z aktami prawnymi Unii Europejskiej, które zobowiązują producentów do sporządzania tłumaczenia dokumentacji produktów w językach krajów, w których mają być one sprzedawane.

W relatywnie dużych korpusach równoległych, a do takich należy czesko-polski InterCorp, użytkownik może poszukiwać ekwiwalentów przekładowych za pomocą Treq w poszczególnych kolekcjach (funkcja *omezit na*). Możliwość ograniczenia poszukiwań do określonych kolekcji ma istotne znaczenie dla trafności pozyskiwanych ekwiwalentów przekładowych. W odróżnieniu od programu Kontext[6], który nie sugeruje ekwiwalentów przekładowych, a które użytkownik

[6] Podstawową aplikacją internetową wykorzystywaną do pracy z zasobami Narodowego Korpusu Języka Czeskiego jest KonText, będącą udoskonaloną i rozszerzoną wersją wcześniejszej NoSketchEngine.

określa sam, Treq wskazuje po kolei te, które posiadają najwyższą częstość użycia. Listę ewentualnych translatów uporządkowaną malejąco otwiera jednostka mająca najwięcej poświadczeń, zamyka – notowana najrzadziej.

Każdy wybór użytkownika Treq może mieć wpływ na liczbę i jakość pozyskanych poświadczeń. Pierwszy z nich to zaznaczenie funkcji *lemmaty (*czes. *lemmata)*, co umożliwia pozyskanie większej liczby poświadczeń poprzez wyszukanie wszystkich wyrazów gramatycznych określonego leksemu[7].

Dla ilustracji posłużmy się przykładem polskiej jednostki *ciacho*, dla którego postaramy się ustanowić ekwiwalenty przekładowe za pomocą Treq. Analiza tej jednostki ewokuje wiele problemów, które pojawiają się podczas analizy korpusowej, stąd jego wybór.

Odwołanie się do drukowanych lub elektronicznych słowników jednojęzycznych czy przekładowych, czyli źródeł systemowych, przynosi zróżnicowane rezultaty.

Inny słownik języka polskiego notuje leksem *ciacho* jako zgrubienie od *ciastko* i odnotowuje jego potoczność (ISJP 2000, s. 183)

Precyzyjne eksplikacje znaczeń leksemu *ciacho* podaje elektroniczny *Wielki słownik języka polskiego*, którego główną bazą materiałową jest Narodowy Korpus Języka Polskiego (NKJP). Pierwsze znaczenie to ‘ekspresywnie o ciastku’, drugie – ‘mężczyzna bardzo atrakcyjny fizycznie’. Oba opatrzone są w słowniku kwalifikatorami *potoczne (pot.)*.

Polsko-český slovnik Karela Olivy dla polskiego zgrubienia *ciacho* podaje trzy czeskie ekwiwalenty: *dort*, *řez* i *cukrářský kousek* (PČS 1999, s. 136).

Polsko-czeski słownik Lingea 5 podaje parę ekwiwalentów *ciacho – fešák* (https://www.dict.com/czesko-polski/ciacho), czyli nie notuje

[7] Ponieważ termin *leksem* traktujemy jako jednostkę systemu językowego, to staramy się jego unikać w stosunku do analizowanych jednostek tekstowych, tworzących pary przekładowe.

znaczenia zgrubienia *ciastko*, a wskazuje na osobę posiadającą określone cechy.

Uzasadnienie takiego wyboru znajdujemy w *Słowniku nieliterackiego języka czeskiego*, który notuje dwa znaczenia leksemu *fešák*: 1. ‘elegantní, dobře vypadající a oblečený muž’, 2. *vulg.* ‘mužské přirození’ (SNČ 2009, s. 128).

Jak widać, dwa spośród czterech słowników uwzględniają znaczenie leksemu *ciacho* w odniesieniu do osoby. Nieodnotowanie tego znaczenia w dwu pracach leksykograficznych tu przywołanych jest usprawiedliwione. Jest to znaczenie, które wyodrębniło się w latach dziewięćdziesiątych minionego stulecia, a pierwsze jego poświadczenie w NKJP pochodzi z 1999 roku. Jego źródłem jest krakowski „Dziennik Polski”, który przytacza definicję leksemu *ciacho* w znaczeniu ‘dodatnio o chłopaku (w ocenie dziewcząt)’ ze *Słownika współczesnej gwary uczniowskiej Anno Domini 1999* (http://nkjp.uni.lodz.Pl/ParagraphMetadata?pid=af0b84430dcead8c4d037923ce9c8d8d&match_start=1&match_end=7&wynik=114#the_match). Wszystkie wcześniejsze, nieliczne zresztą, poświadczenia korpusowe tej jednostki (pierwsze z 1993 roku) w NKJP to zgrubienia od *ciastko.* Można więc stwierdzić, że jednostka *ciacho* w podanym znaczeniu odbyła drogę z gwary młodzieżowej do polszczyzny potocznej, potwierdzając tym samym silny wpływ tej gwary na inne odmiany polszczyzny.

W podanej w *Polsko-czeskim słowniku Lingea 5* parze ekwiwalentów przekładowych *ciacho – fešák* trudno nie dostrzec braku pełnej symetrii znaczeniowej. Potwierdzają to definicje z przywołanych tu słowników jednojęzycznych; w przypadku polskiej jednostki podkreślana jest atrakcyjność fizyczna mężczyzny, w przypadku jego czeskiego odpowiednika – jego elegancja i atrakcyjny wygląd.

Przejdźmy teraz od źródeł systemowych do tekstowych, czyli do programu Treq, i spróbujmy ustalić ekwiwalent lub ekwiwalenty dla analizowanej jednostki w odniesieniu do osoby. Treq dla zapytania *ciacho* bez zaznaczenia funkcji *lemmaty* przynosi następujące infor-

macje (ustawienia w kolejności: výchozí jazyk: polština, cílový jazyk: čeština, omezit na: kolekce 5)[8]:

Tabela 1. Treq dla zapytania *ciacho*

Liczba poświadczeń (Frekvence)	Udział procentowy (Procenta)	Polski transland (Polština)	Czeski ekwiwalent (Čeština)
8	30,8	ciacho	sexy
2	7,7	ciacho	krásnej
1	3,8	ciacho	výnikající
1	3,8	ciacho	třída
1	3,8	ciacho	Sekne
1	3,8	ciacho	Ni
1	3,8	ciacho	kus
1	3,8	ciacho	koláček
1	3,8	ciacho	fešák
1	3,8	ciacho	dej
1	3,8	ciacho	výborně
1	3,8	ciacho	sluší
1	3,8	ciacho	Randit
1	3,8	ciacho	moučník
1	3,8	ciacho	krasavec
1	3,8	ciacho	Keksi
1	3,8	ciacho	extrémně
1	3,8	ciacho	báječný

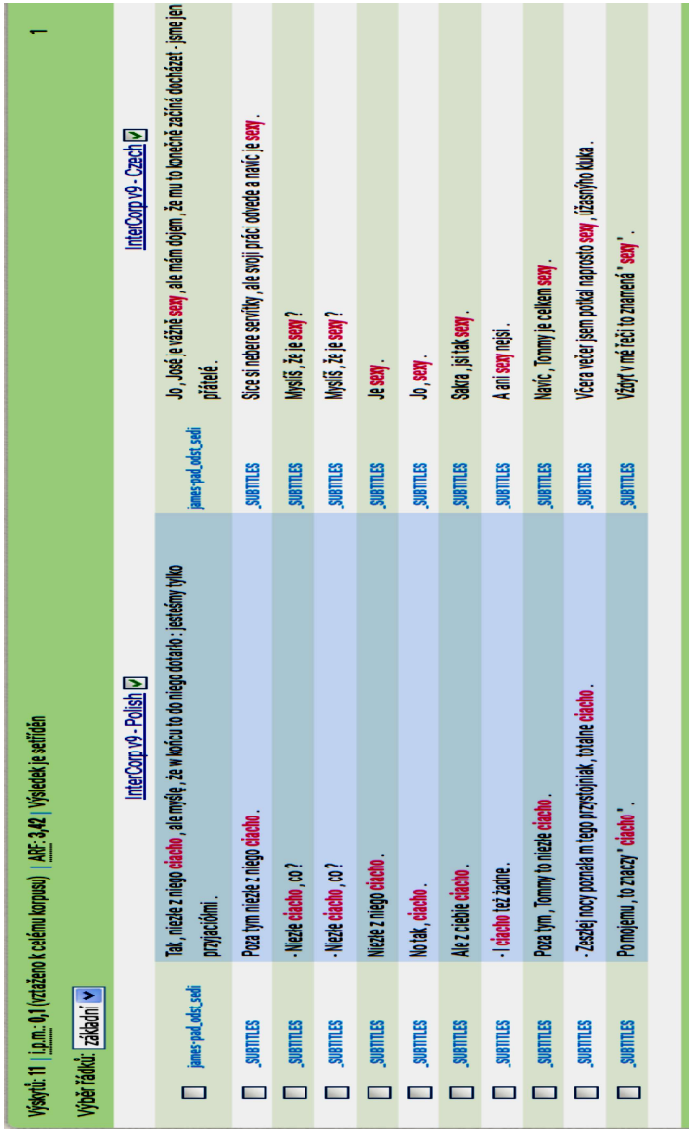
[8] Ze względu na brak miejsca nie zawsze prezentuję listy ze wszystkimi poświadczeniami, których może być niekiedy kilkadziesiąt i więcej.

Zaznaczenie funkcji *lemmaty* pozwala na uzyskanie ponad dwa razy większej liczby poświadczeń – 56 (*sexy* – 12, *on* – 2, *fešák* – 2, úžasný – 2, *kus* – 2, pozostałych 36 – 1). Istotne są źródła pozyskania poświadczeń, które użytkownik w każdej chwili może sprawdzić. Prawie wszystkie poświadczenia (odpowiednio 25 z 26 (bez funkcji *lemmaty* i 49 z 56 z funkcją *lemmaty*) pochodzą z kolekcji napisów filmowych, czyli z dialogów w sytuacji nieoficjalnej, potwierdzając tym samym ich przynależność stylistyczną do potocznej odmiany polszczyzny.

Uzyskane poświadczenia są rezultatem automatycznej analizy korpusowej, a ta – jak wiadomo – może być obarczona sporymi błędami. Mogą one wynikać zarówno z błędnej anotacji, jak i ze zjawiska homonimii. Tak więc, mając do czynienia z programem Treq, który nie rozróżnia znaczeń jednostek polisemicznych, należy się liczyć z poświadczeniami tego, czego w istocie nie szukamy. W przypadku zaznaczenia funkcji *lemmaty* np. jedno z poświadczeń (*ciach*) nie jest formą dopełniacza liczby mnogiej rzeczownika *ciacha*, a wykrzyknikiem *ciach.* Oczywiście, podobne błędy wymuszają ręczną analizę kontekstów, będącą koniecznym uzupełnieniem analizy automatycznej. Oto paralelne konkordancje pozyskane po przejściu z Treq do zasobów InterCorp v9 dla sugerowanej pary ekwiwalentów *ciacho* – *sexy* bez zaznaczania funkcji *lemmaty* (por. tabela 2).

Ręczna analiza kontekstów przynosi wiele istotnych informacji. Po pierwsze, potwierdza mankamenty analizy automatycznej – liczba poświadczeń (8) nie zgadza się z liczbą wyświetlonych kontekstów (11), jeden kontekst jest powtórzony. Po drugie, ani razu zestawiana para *ciacho* – *sexy* nie jest wynikiem tłumaczenia bezpośredniego. Językiem oryginału jest angielszczyzna, a jedyny wyjątek stanowi tłumaczenie z języka duńskiego. W praktyce oznacza to, że we wszystkich przypadkach mamy polsko-czeską parę ekwiwalentów zestawioną z przekładów, czyli translatów angielskich i duńskich translandów. Po trzecie, praktycznie wszystkie poświadczenia pochodzą z napisów filmowych, jeden tylko z literatury pięknej, a ściślej powieści *Pięćdziesiąt twarzy Greya* brytyjskiej pisarki E. L. James.

Tabela 2. Konkordancje dla pary ekwiwalentów *ciacho* – *sexy* bez zaznaczania funkcji *lemmaty*, pozyskane po przejściu z Treq do zasobów InterCorp v9



Wskazywany jako ewentualny ekwiwalent *fešák* w rozszerzonym przeszukiwaniu z funkcją *lemmaty* ma tylko 3 poświadczenia, będące przekładami z angielskiego (2) i chińskiego (1). Tak znikoma liczba poświadczeń daje podstawy do upatrywania w *fešák* ekwiwalentu jednostki *ciacho* w odniesieniu do mężczyzny w określonych kontekstach użycia, ale nie daje pewności.

Możliwość zmiany kierunku tłumaczenia pozostaje z pewnością zaletą aplikacji Treq. Można w ten sposób zweryfikować prawidłowość ustanowionych par ekwiwalentów, a oprócz tego wskazać nowe. Z analizy danych korpusowych wynika, że polski rzeczownik *ciacho* w znaczeniu ‘mężczyzna atrakcyjny fizycznie’ posiada w czeszczyźnie dwa ekwiwalenty *sexy* i *fešák*. Oczywiście, nie wykluczając innych możliwych ekwiwalentów tekstowych. Przy zmianie kierunku tłumaczenia na czesko-polski, Treq przedstawia następujące dane dla *sexy* (ustawienia w kolejności: výchozí jazyk: čeština, cílový jazyk: polština, omezit na: kolekce 5):

Tabela 3. Treq dla zapytania *sexy*

Liczba poświadczeń	Udział procentowy	Czeski transland	Polski ekwiwalent
145	17.1	sexy	sexy
96	11.3	sexy	seksowny
88	10.4	sexy	seksowne
86	10.2	sexy	seksowna
37	4.4	sexy	seksownie
31	3.7	sexy	seksowna
30	3.5	sexy	qoraca
25	3.0	sexy	seksownego
14	1.7	sexy	qorący
12	1.4	sexy	seksownym
11	1.3	sexy	seksy
11	1.3	sexy	podniecające
10	1.2	sexy	qorace
10	1.2	sexy	sexi

9	1.1	sexy	laska
9	1.1	sexy	seksownej
8	0.9	sexy	seksi
8	0.9	sexy	ciacho

Analiza danych korpusowych potwierdza, że w przypadku polskich odpowiedników czeskiego *sexy* najczęściej wskazywany jest polski przymiotnik (używany również jako przysłówek) *sexy* i jego spolszczone formy. Praktycznie wszystkie poświadczenia pochodzą z platformy OpenSubtitles i są przekładami z języka angielskiego, tylko trzy to tłumaczenia bezpośrednie: 1. polsko-czeskie (kolekcja PressEurope) i 2 czesko-polskie (tłumaczenia powieści Michala Vievegha *Zapisywacze ojcowskiej miłości* i *Sprawa niewiernej Klary*). Znamienne, że w żadnym z tych trzech tłumaczeń ani jeden nie odnosi się do osoby (*Sláva je* ***sexy***, *přízvuk je* ***sexy***, *EU není moc* ***sexy***). Przytoczone poświadczenia dobrze ilustrują mankamenty amatorskich napisów filmowych, tj. błędy tłumaczeniowe i językowe. Niektóre z nich, jak można dostrzec, mają wysoką frekwencję. Ich wychwycenie zależy od kompetencji samego użytkownika Treq, który powinien poddać je wnikliwej analizie.

Wyniki polskich ekwiwalentów dla *fešák* prezentuje tabela 4 (ustawienia w kolejności: výchozí jazyk: čeština, cílový jazyk: polština, omezit na: kolekce 5).

Tabela 4. Treq dla zapytania *fešák* bez żadnej dodatkowej funkcji

Liczba poświadczeń	Udział procentowy	Czeski transland	Polski ekwiwalent
59	34.1	fešák	przystojny
9	5.2	fešák	Przystojny
8	4.6	fešák	Przystojniak
7	4.0	fešák	przystojniak
6	3.5	fešák	słodki
5	2.9	fešák	przystojniaczek
4	23	fešák	niezły
4	2.3	fešák	ładny

3	1.7	fešák	miły
3	1.7	fešák	piękny
2	1.2	fešák	boski
2	1.2	fešák	czarujący
2	1.2	fešák	fantastyczni
2	1.2	fešák	miłym
2	12	fešák	sztuka
2	12	fešák	wyładniał
2	1.2	fešák	śliczny

Analiza wskazywanych przez Treq ekwiwalentów przekładowych w zasadzie nie uwzględnia jednostki *ciacho* (1 poświadczenie, 2 konteksty użycia w OpenSubtitles: *Je to fešák. – Ale z niego ciacho!*; *Jsi fešák. – Fajne z ciebie ciacho.)*. Zwracają uwagę natomiast pary *fešák – przystojniak* (wszystkie z napisów filmowych) oraz *fešák – przystojny (mężczyzna/ facet)* (15 kontekstów użycia w literaturze, jedno tłumaczenie bezpośrednie polsko-czeskie).

W przypadku jednostki *fešák* istotne znaczenie ma zapytanie z funkcją *lemmaty*. Wyniki prezentauje tabela 5 (ustawienia w kolejności: výchozí jazyk: čeština, cílový jazyk: polština, omezit na: kolekce 5, lemmat włączony).

Tabela 5. Treq dla zapytania *fešák* z funkcją *lemmaty*

Liczba poświadczeń	Udział procentowy	Czeski transland	Polski ekwiwalent
91	22,5	fešák	przystojny
45	11.1	fešák	przystojniak
25	6.2	fešák	buźka
22	5.4	fešák	przystojniaczek
10	2.5	fešák	miły
9	22	fešák	ładny
8	2.0	fešák	wspaniały
7	1.7	fešák	słodki
7	1.7	fešák	piękny

6	1.5	fešák	blood
5	1.2	fešák	niezły
5	1.2	fešák	śliczny
4	1.0	fešák	towar
4	1.0	fešák	chłopiec
4	1.0	fešák	pilocik
4	1.0	fešák	wyglądać
4	1.0	fešák	chłoptaś

To, co może zaskakiwać, to pojawienie się na trzeciej pozycji jednostki *buźka* z 25 poświadczeniami i ponad 6-procentowym udziałem w ogólnej liczbie ekwiwalentów. Po wejściu na konteksty użycia (27) okazuje się, że wszystkie poświadczenia (z wyjątkiem jednego pisanego małą literą, będącego wynikiem wątpliwego tłumaczenia z angielskiego) odnoszą się do tego samego źródła – serialu *Drużyna A* (*The A-Team*) i jednego z jej bohaterów – Buźki. Bez zaznaczenia funkcji *lemmaty* czeski zwrot adresatywny *Fešáku* nie jest znajdowany przez aplikację Treq. Dla wprowadzonej jednostki *Fešák* pisanej wielką literą program bezbłędnie wyszukuje 22 poświadczenia pary adresatywów *Fešáku – Buźka*.

Dla przeprowadzonej analizy fakty te mają spore znaczenie, ponieważ uświadamiają po raz kolejny problem analizy automatycznej – homonimii, a ściślej nieodróżnienia nazwy własnej (*Buźka*) od nazwy pospolitej (*buźka*). Ponadto, jak widać, wyszukiwanie odpowiedników przekładowych dla jednej jednostki niejako przy okazji umożliwia ustalanie innych par przekładowych.

Podsumujmy. Analizując pozyskany materiał językowy i biorąc pod uwagę trzy kryteria ekwiwalencji – semantyczne, stylistyczne i pragmatyczne, można ustalić następujące pary przekładowe dla *ciacho* w znaczeniu ‘mężczyzna bardzo atrakcyjny fizycznie’:

CIACHO – SEXY/FEŠÁK
PRZYSTOJNIAK/PRZYSTOJNY FACET – FEŠÁK

Wydaje się, że para *ciacho – sexy* adekwatnie oddaje semantykę zestawianych polskich i czeskich ekwiwalentów – seksualność. Obie

jednostki należą do tego samego rejestru stylistycznego – odmiany potocznej. Jedyna częściowa niezgodność między polskim *ciacho* a czeskim *sexy* dotyczy sfery pragmatyki – czeska jednostka może być zarówno określeniem mężczyzny, jak i kobiety.

Druga para przekładowa pojawiła w trakcie ustalania ekwiwalentów dla jednostki *ciacho*. Choć Treq wskazuje jednoznacznie na ekwiwalencję pary *przystojniak / przystojny facet – fešák*, to należy odnotować brak pełnej symetrii semantycznej między tymi jednostkami. Otóż znaczenie polskiego leksemu *przystojniak*[9] odnosi się do atrakcyjnej fizyczności mężczyzny (zwykle z perspektywy kobiety), podczas gdy czeski *fešák* określa mężczyznę dobrze ubranego, dbającego o swój wygląd. W takiej sytuacji jako czeski ekwiwalent można byłoby wskazać jednostkę *krasavec*[10], choć ten w odróżnieniu od potocyzmu *fešák* nie spełnia kryterium stylistycznego – należy do ogólnej odmiany czeszczyzny. Jak z powyższego wynika, kryterium statystyczne nie może być jedyne we wskazywaniu par ekwiwalentów, może być jednym z wielu.

W tym miejscu należy też wspomnieć o innych funkcjach Treq, których użytkownik może użyć w poszukiwaniu ekwiwalentów przekładowych.

Funkcja *regulární* umożliwia bardziej zaawansowane formułowanie zapytań. Ta popularna funkcja korpusowa polega na wykorzystywaniu w zapytaniach tzw. wyrażeń regularnych, czyli symboli, służących do tworzenia sekwencji w przypadku wyszukiwania określonych zbiorów wyrazów. Wyrażenia regularne mogą się składać wyłącznie ze znaków specjalnych, być kombinacją znaków specjalnych i alfanumerycznych lub zawierać tylko znaki alfanumeryczne. Funkcja ta jest bardzo przydatna w określaniu ekwiwalencji zarówno jednostek jednowyrazowych, przez nas tu analizowanych, jak i wielowyrazowych, które zasługują na odrębną analizę.

Najbardziej uniwersalnym znakiem jest kropka (.), która może zastąpić dowolną literę. Dla przykładu sekwencja trzech kropek z rzędu umożliwi znalezienie jednostek trzyliterowych, czterech kropek – czteroliterowych itd. W wyszukiwaniu określonych form może okazać się przydatna gwiazdka (*), która zastępuje dowolny ciąg znaków (zero i więcej). Znak plus (+) pełni podobną funkcję, różniąc się od gwiazdki tym, że zastępuje co najmniej jeden znak lub więcej (jeden i więcej).

Zastosowanie jednego, drugiego czy trzeciego znaku może przynieść różną liczbę poświadczeń. Ważne jest prawidłowe zbudowanie zapytania. Znaki o podobnych funkcjach mogą przynosić różną liczbę poświadczeń[11].

Funkcja (A=a) pozwala na nierozróżnianie wielkich i małych liter, a więc jej zaznaczenie umożliwi znalezienie np. jednostek Ciacho, ciacho, CIACHO.

Mogą być również użyte inne znaki, pozwalające na wyodrębnienie wariantów głównie jednostek wielowyrazowych: nawiasy okrągłe, nawiasy okrągłe, nawiasy kwadratowe, odwrócone ukośniki, linie pionowe, daszki, myślniki, litery, cyfry. Wszystkie one mogą być użyte w zapytaniu.

Jak widać, możliwości, jakie Treq stwarza użytkownikowi, nie są wcale małe. Można nawet stwierdzić, że tym większe, im większe są jego wiedza o języku, znajomość narzędzia Treq oraz zaangażowanie w poszukiwanie optymalnych rozwiązań przekładowych. Przeprowa-

[9] Definicja leksemu *przystojniak* w *Wielkim slowniku języka polskiego* jest następująca '*pot.* przystojny mężczyzna'.

[10] Narzędzie Treq przy wyszukiwaniu ekwiwalentów przekładowych dla polskiego leksemu *przestojniak* przy zaznaczonej funkcji *lemmaty* wskazuje na pierwszym miejscu *fešák* (45 poświadczeń), na drugim – *krasavec* (43 poświadczenia), na trzecim – *hezoun* (16 poświadczeń). Pod wieloma względami leksemy *przystojniak* i *krasavec* można uznać za parę przekładową. To samo można powiedzieć o leksemie *hezoun*, który podobnie jak *przystojniak* może być użyty z odcieniem ironii czy lekceważenia. Co prawda, *hezoun* jest opatrzony kwalifikatorem *ekspresywny* (*expresivní výraz*) (SSJČ), a polski *przystojniak* kwalifikatorem *potoczne* (WSJP), to jednak konteksty użycia obu wyrazów wskazują jednoznacznie na ich ekwiwalencję.

[11] Omówienie funkcji poszczególnych znaków wraz z przykładami użycia znajdziemy na stronie https://wiki.korpus.cz/doku.php/kurz:regularni_vyrazy.

dzona analiza pokazuje, że Treq może być relatywnie skutecznym narzędziem wspomagającym proces tłumaczenia. Narzędzie Treq dobrze sprawdza się w wyszukiwaniu jednostek jednowyrazowych. Proces ten przebiega szybko i sprawnie, w praktyce nie wymaga od użytkownika specjalistycznego przygotowania, np. znajomości wyrażeń regularnych. Uzyskane ekwiwalenty przekładowe z reguły są trafne, a ponadto są (zwykle jest ich więcej niż jeden) automatyczne sklasyfikowane zgodnie z kryterium frekwencyjnym, które w aplikacji Treq jest kryterium podstawowym.

Oceniając wartość aplikacji Treq dla procesu ustalania polsko-czeskich par przekładowych, można sformułować kilka ogólnych wniosków.

Aplikacja Treq dobrze sprawdza się w wyszukiwaniu ekwiwalentów przekładowych jednostek jednowyrazowych. Pozyskane tą drogą dane korpusowe zwykle trafnie wskazują parę przekładową. Dużą zaletę posiada sporządzana przez Treq lista frekwencyjna, uwzględniająca inne, rzadziej używane, a przydatne w określonych kontekstach alternatywne ekwiwalenty. Pokaźna liczba poświadczeń, będąca pochodną dużych zasobów polsko-czeskiego korpusu równoległego InterCorp v9 gwarantuje dodatkowo adekwatność ustalanych par przekładowych. Potwierdza się teza, że automatyczne wyszukiwanie ekwiwalentów przekładowych obarczone jest ryzykiem wskazywania jako ekwiwalenty jednostek nietrafnych. Przyczyny tego zjawiska mogą być różne, często natury technicznej związanej z przetwarzaniem zgromadzonego materiału. Najczęściej jednak można wskazać trzy: wadliwą anotację, zjawisko homonimii i nieadekwatne tłumaczenie (zwłaszcza w napisach z platformy Opensubtitles). Rzecz jasna, nie należy zapominać o porównywanych ze sobą polskich i czeskich przekładach zwykle angielskiego oryginału. Okoliczność ta z pewnością nie sprzyja pozyskiwaniu adekwatnych wyników. Nie powinny zatem zaskakiwać wskazywane przez Treq propozycje błędne, które są nieodłącznym elementem rzeczywistości korpusu równoległego.

Przeprowadzona przy okazji omówienia funkcjonalności Treq analiza pokazuje, że pomimo swoich mankamentów narzędzie to pozwala skutecznie definiować znaczenie (znaczenia) analizowanych jednostek, a oprócz tego wskazać to, co w przekładzie jest najistotniejsze – aby ustalone pary przekładowe posiadały analogiczną wartość komunikacyjną.

Skróty

ISJP	*Inny słownik języka polskiego*, (Ed.) M. Bańko, t. 1. Warszawa: Wydawnictwo Naukowe PWN, 2000.
NKJP	Narodowy Korpus Języka Polskiego. Online: http://www.nkjp.pl [dostęp: 02.02.2020]
PČS	Oliva K. (a kol.): *Polsko-český slovník*, T. 1–2. Praha: Academia, 1999.
SNČ	*Slovník nespisovné češtiny*, Praha: *Maxdorf, 2009.*
SSJČ	*Slovník spisovného jazyka českého*. Online: https://ssjc.ujc.cas.cz [dostęp: 02.02.2020]
VSPČČP	*Velký slovník polsko-český i česko-polský* (*elektronický slovník*), Lexicon 5. Brno: LINGEA, 2010.
WSJP	*Wielki słownik języka polskiego*. Online: https://www.wsjp.pl [dostęp: 02.02.2020]

Literatura

Čermák F., 2017, *Korpus a korpusová lingvistika*, Praha: Karolinum.

Och F. J., Ney H., 2003, *A systematic comparison of various statistical alignment models*. „*Computational Linguistics*" 1:29, s. 19–51.

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