ČASOPIS PRO MODERNÍ FILOLOGII 97, 2015, Č. 2, S. 157–168

Jak najít optimální překlad polysémních sloves — porovnání metod formální analýzy paralelních textů[1]

Elżbieta Kaczmarska (Warszawa) — Alexandr Rosen (Praha)

IN SEARCH OF THE OPTIMAL TRANSLATION OF POLYSEMOUS VERBS — COMPARING METHODS OF FORMAL ANALYSIS OF PARALLEL TEXTS

Our goal is to identify factors that influence the choice of equivalents of 'psych' verbs when translating between typologically close languages such as Polish and Czech. Using the example of the Czech verb *toužit* 'to yearn, to desire' we show that these verbs may be perceived differently by native speakers of Polish and Czech — as ambiguous or unambiguous. Translation of such verbs is equally challenging. We start with the hypothesis that the choice of an equivalent is determined primarily by syntactico-semantic properties of the source lexeme, especially by its valency. Based on the analysis of lexemes and their arguments in parallel texts we identify regularities and preferences for the choice of an equivalent. Manual analysis is complemented by an automatically extracted bilingual glossary with frequencies. The results show that valency is an important, but not the only factor.

KEYWORDS
Czech, Polish, 'psych' verbs, valency, parallel corpus

KLÍČOVÁ SLOVA
čeština, polština, slovesa duševních stavů, valence, paralelní korpus

1. ÚVOD

Čeština a polština vykazují při srovnání řadu jevů charakteristických pro typologicky velmi blízké jazyky, např. vysokou frekvenci tzv. zrádných slov (viz např. Lotko, 1992). Časté jsou však i jevy typické spíše pro jazyky vzdálenější, např. problémy s identifikací vhodného lexikálního ekvivalentu.[2] Příkladem mohou být pojmy vyjadřující duševní stavy, např. *mít rád*, *být líto* nebo *toužit*. Slovesa označující city a emoce jsou z tohoto pohledu zvlášť problematická pro svou víceznačnou, náznakovou a subjektivní povahu. Ekvivalent se pak hledá velmi obtížně a část původního významu se při překladu ztrácí. Někdy ani není možné význam výrazu v cílovém jazyce vyjádřit, protože odpovídající pojem chybí (Kaczmarska — Rosen, 2014b). Lze ho opsat jinými slovy (např. *kilkanaście* — „deset až dvacet") nebo aproximovat více ekvivalenty, ale

1 Za inspiraci a obětavou pomoc s využitím programových nástrojů děkujeme J. Hanovi, B. Vidové, O. Bojarovi, D. Marečkovi, H. Skoumalové a T. Jelínkovi. Práce na tomto projektu byla částečně podpořena z grantu MŠMT Český národní korpus, č. LM2011023.

2 V této práci se nezabýváme definicí ekvivalence a ekvivalentu. Viz např. Baker (1992), Catford (1965), Dąmbska-Prokop (2000), Hejwowski (2009), Koller (1995) či Nida (1995).

žádný z nich samostatně ani všechny jako celek nemusí pokrýt přesně stejné sémantické pole.[3] Rodilí mluvčí češtiny takové výrazy jako polysémní většinou nevnímají,[4] existence příslušného výrazu v jiném jazyce a jeho absence v jazyce vlastním však bývá pro polské mluvčí důvodem neporozumění obsahu sdělení.

Tradiční slovníky nabízejí jen omezený počet ekvivalentů, většinou bez příkladů užití. Například pro *toužit* uvádí česko-polský slovník[5] tři významově odlišné ekvivalenty: *tęsknić*,[6] *marzyć*[7] a *pragnąć*.[8] Bez dalších informací (valence, kolokace) je pro uživatele nemožné vybrat ekvivalent, který do kontextu zapadne. Někdy však ani kontext nepomůže: vyznání *mám tě rád* lze přeložit jako *lubię cię*[9] nebo jako *kocham cię*[10] s podstatně odlišným významem.

Cílem této práce je ověřit hypotézu, že vhodné ekvivalenty pro slovesa označující psychické stavy lze stanovit na základě formálně uchopitelné syntakticko-sémantické analýzy argumentů těchto sloves. Postupujeme tak, že v kontextu českého lexému zkoumáme faktory, které vedou k volbě polského ekvivalentu. Vycházíme přitom z ručně roztříděných paralelních konkordancí lexému *toužit* a jeho ekvivalentů jako reprezentativního příkladu slovesa s výraznou polysémií (alespoň z hlediska polských mluvčích) a s několika významově odlišnými ekvivalenty (viz též Kaczmarska — Rosen, 2013). Výsledky doplňujeme a ověřujeme seznamem ekvivalentů s údaji o frekvenci, excerpovaných automaticky z paralelních textů na základě zarovnání po slovech. Vzhledem k tomu, že původní hypotézu lze potvrdit jen částečně a automaticky excerpovaný dvoujazyčný glosář neumožňuje predikci optimálního ekvivalentu na základě kontextu, rozšiřujeme analýzu na další faktory s využitím stochastického klasifikátoru a syntakticky strukturovaného kontextu v paralelních textech. Podrobnější popis a evaluace jsou však mimo rámec článku.

3 Podle Lewandowské-Tomaszczyk (1984, 2013) stoprocentní ekvivalenty neexistují; je ale možné najít množinu sémanticky blízkých jednotek (*a cluster of equivalents*) a využít je při překladu.

4 Např. SSJČ (Havránek et al., 2011) definuje *mít rád* v bodu 2 u hesla *rád* takto: *pociťovat k někomu náklonnost, lásku*; *milovat*[2,3]; *mít v oblibě, milovat*[3]. Odkazy na *milovat* sice vylučují „mileneckou lásku" (*milovat*[1]), ale z definice i z příkladů (*mají se rádi a budou se brát*) je zřejmé, že významové pole výrazu je široké a značně homogenní.

5 Viz Siatkowski — Basaj (2002), který obsahuje 53 tisíc hesel a 28 tisíc výrazů; standardní polsko-český slovník (Oliva, 1994) obsahuje 80 tisíc hesel.

6 Překlad definic podle internetové verze výkladového slovníku *Słownik Języka Polskiego* — http://sjp.pwn.pl: 1) cítit žal, být smutný kvůli nepřítomnosti jiné osoby, absenci kontaktu s někým nebo s něčím, 2) silně toužit (*pragnąć*) získat něco, dosáhnout něčeho (http://sjp.pwn.pl/szukaj/tęsknić.html)

7 1) představovat si to, po čem se touží (*pragnąć*), přemítat o příjemných věcech, často nereálných, 2) velmi silně po něčem toužit (*pragnąć*), 3) *arch.* snít (http://sjp.pwn.pl/szukaj/marzyć.html)

8 1) velmi něco chtít, 2) dychtit (*pożądać*) po někom, 3) chtít něco říct, vysvětlit (http://sjp.pwn.pl/szukaj/pragnąć.html)

9 1) chovat k někomu sympatie, 2) nacházet v něčem zalíbení, 3) o rostlinách, zvířatech, věcech: vyžadovat, potřebovat něco (http://sjp.pwn.pl/szukaj/lubić.html)

10 obdařovat někoho pocitem lásky, někoho/něco hodně mít rád; také: chovat vůči osobě opačného pohlaví vřelé city spojené s erotickou touhou (http://sjp.pwn.pl/szukaj/kochać.html)

V části 2 popisujeme ruční analýzu korpusových konkordancí, jejíž souhrn pak představujeme v části 3. V části 4 se zabýváme excerpcí dvoujazyčného glosáře a jeho srovnáním s předchozími výsledky. Poslední část 5 se pak věnuje diskusi a perspektivám, včetně možností, jak využít metody strojového učení k nalezení kontextově nejvhodnějšího ekvivalentu obtížně přeložitelných lexémů.

2. RUČNÍ ANALÝZA

Je valenční struktura českých sloves duševních stavů významným faktorem při volbě jejich polských ekvivalentů? Zde se na tuto otázku pokoušíme odpovědět metodou ruční analýzy paralelních konkordancí.[11]

Data pocházejí z 5. vydání paralelního korpusu *InterCorp*.[12] Jeho jádro tvoří originály a překlady beletristických textů s ručně zkontrolovaným zarovnáním po větách. Zbytek je z větší části právnická a publicistická literatura, zpracovaná automaticky. Zde pracujeme jen s česko-polskou beletrií, pro ruční analýzu si z ní vybíráme jen původní české texty v rozsahu 1,8 milionu slov.

Pilotní studie slovesa *toužit* (Kaczmarska — Rosen, 2013) byla založena na předpokladu, že pro některé významy může být ekvivalent určen konvergencí valenčních požadavků.[13] Při syntaktické a sémantické analýze argumentů českého slovesa a jeho polských ekvivalentů jsme postupovali takto: (i) vyhledali jsme překlad českého slovesa v zarovnaných polských segmentech, (ii) určili jsme počet a typy jeho argumentů (např. substantivum odkazující na lidskou bytost, abstraktní nebo konkrétní entitu) a (iii) jeho morfosyntaktické vlastnosti (prostý nebo předložkový pád, infinitiv, vedlejší věta). Nakonec jsme (iv) prozkoumali argumentovou strukturu ekvivalentů. Očekávali jsme, že výsledky analýzy umožní ve většině případů predikovat nejvhodnější ekvivalent.[14]

11 Valencí v tomto článku rozumíme počet argumentů řízených slovesným predikátem spolu s jejich morfosyntaktickými a sémantickými vlastnostmi (viz např. Dębski, 1982; Daneš — Hlavsa, 1987; Rytel, 1989; Greń — Rytel-Kuc, 1991; Čermáková, 2009; Urbańczyk-Adach, 2001; Kaczmarska, 2001). Jde nám přitom zejména o valenci objektovou. Podrobnější popis syntaktických a sémantických vlastností zkoumaných sloves bude předmětem dalšího výzkumu.

12 Viz Čermák — Rosen (2012) a Kaczmarska — Rosen (2014a). InterCorp je jako část Českého národního korpusu prohledávatelný on-line, viz http://www.korpus.cz/intercorp/. V prosinci 2014 bylo zveřejněno jeho 7. vydání.

13 Vycházíme z předpokladu, že syntaktické chování slovesa závisí do značné míry na jeho významu (Levin, 1993). Bereme také v úvahu možnost, že překladatel preferuje ekvivalent se stejnou nebo podobnou valenční strukturou.

14 Zde je namístě otázka, zda InterCorp je pro takový výzkum dostatečně reprezentativním korpusem, konkrétně zda zastoupení různých valenčních rámců slovesa *toužit* v korpusu InterCorp je srovnatelné s daty z jednojazyčného korpusu. Pro srovnání jsme použili beletristickou část korpusu SYN (120 mil. pozic z celkového počtu 2,7 miliardy, viz http://korpus.cz, přístup dne 27. 11. 2014). Relativní frekvence tvarů slovesa *toužit* je 83,1 ipm (items per million), což je méně než v česko-polské části InterCorpu 5, omezené na české originály, totiž 145.08 ipm. Z toho usuzujeme, že různé typy užití slovesa *toužit* jsou v korpusu InterCorp zastoupeny dostatečně, i když konkrétní poměry mezi frekvencemi různých valenčních rámců mohou být odlišné.

Kromě argumentů lze zkoumat i další větné členy spojené s daným slovesem, např. adverbiale, ale zde se soustředíme na objektové argumenty, u nichž předpokládáme, že budou mít na volbu ekvivalentu největší vliv.[15] Objekty slovesa *toužit* jsme roztřídili do 5 skupin podle kombinace dvou kritérií: sémantické klasifikace a morfosyntaktické realizace (viz tabulka 1).[16] Celkem jsme takto zpracovali všech 246 výskytů slovesa *toužit*. Pro každý typ uvádíme vždy několik příkladů s typem realizace argumentů v polštině a na závěr souhrnnou tabulku, která navíc ukazuje podobně zpracované výskyty slovesa *toužit* v českých překladech z polských originálů.

Typ	**Klasifikace objektu**	**Absolutní frekvence typu**	**Relativní frekvence typu**
toužit po Oh	lidská bytost	38	15 %
toužit po Oa	abstraktum	90	37 %
toužit po Or	konkrétum	35	6 %
toužit Inf	infinitiv	80	33 %
toužit (*po*) S	vedlejší věta *… (po tom,) aby …*	23	9 %
CELKEM		246	100 %

TABULKA 1. Zastoupení valenčních typů slovesa ***toužit*** v originální české beletrii z česko-polské části korpusu *InterCorp* 5

2.1 *TOUŽIT PO* OH

(1) *toužit po* Oh → *pragnąć* Oh (12 výskytů — 32 %)[17]
Jsi krásná, nepřestanu po tobě ***toužit*** *a bát se tvé krásy …*
Jesteś piękna, nigdy nie przestanę cię ***pragnąć*** *i bać się twojej urody …*

(2) *toužit po* Oh → *tęsknić do* S (1 výskyt — 3 %)
Miláčku, já ***netoužím*** *po rodině.*
Kochanie, ja ***nie tęsknię*** *do tego, by założyć rodzinę.*

(3) *toužit po* Oh → *marzyć o* Oh (2 výskyty — 5 %)
Vždycky ***jsem toužila*** *po člověku, který by byl prostý a přímý.*
Zawsze ***marzyłam*** *o człowieku, który był by prosty i bezpośredni.*

15 Tento předpoklad se podařilo nezávisle potvrdit: stochastický klasifikátor vybral ze všech větných členů syntakticky závislých na slovese objektové argumenty jako ty, které o volbě ekvivalentu rozhodují nejvíce (Kaczmarska et al., v tisku).

16 Primární syntaktická klasifikace na základě valenčního slovníku (Lopatková et al., 2014), která v daném případě není dostatečně vypovídající, byla prohloubena o sémantické třídy objektu. Taková klasifikace má za následek, že někdy je třídu objektu obtížné určit jednoznačně, např. výrazy jako *rodina*, *domov*, části těla nebo *láska* lze interpretovat při ruční analýze podle kontextu alespoň dvěma způsoby.

17 Procenta udávají podíl na celkovém počtu dokladů daného českého valenčního typu.

2.2 *TOUŽIT PO* OA (90 VÝSKYTŮ)

(4) *toužit po* Oa → *pragnąć* Oa (29 výskytů — 32 %)
Ale zatím chce, abych život snášel a po smrti ***toužil****.*
A tymczasem chce, bym życie znosił, a śmierci ***pragnął****.*

(5) *toužit po* Oa → *tęsknić do* Oa (11 výskytů — 12 %)
Já ***toužím*** *po lásce.*
Ja ***tęsknię*** *do miłości.*

(6) *toužit po* Oa → *tęsknić za* Oa (7 výskytů — 8 %)
Netoužím *po tomhle slizkém bratrství.*
Nie tęsknię *za takim oślizłym braterstwem.*

(7) *toužit po* Oa → *marzyć o* Oa (20 výskytů — 22 %)
Byli jsme unaveni, promočeni a ***toužili jsme*** *po odpočinku.*
Byliśmy zmęczeni, przemoczeni i ***marzyliśmy*** *o odpoczynku.*

(8) *toužit po* Oa → *pożądać* Oa (5 výskytů — 6 %)
Ale po té slasti Bernard ***netoužil****.*
Ale Bernard takiej rozkoszy ***nie pożądał****.*

2.3 *TOUŽIT PO* OR

(9) *toužit po* Or → *tęsknić za* Or (2 výskyty — 13 %)
Celý život ***jsem toužila*** *po skutečném domově.*
Całe życie ***tęskniłam*** *za prawdziwym domem.*

(10) *toužit po* Or → *marzyć o* Or (5 výskytů — 33 %)
Mladý muž ***touží*** *po vlastním divadle.*
Młody mężczyzna ***marzył*** *o własnym teatrze.*

(11) *toužit po* Or → *pragnąć* Or (3 výskyty — 20 %)
Mé patro, vyprahlé po noci zpola probdělé a zpola neklidně prosněné,
toužilo *po jejím vřelém a mrazivě vonném doušku.*
Moje podniebienie, wyschnięte po nocy na wpół przemarzonej
i na wpół prześnionej, ***pragnęło*** *jej gorącego, orzeźwiająco wonnego łyku.*

2.4 *TOUŽIT* INF

(12) *toužit* Inf → *chcieć* Inf (20 výskytů — 25 %)
Toužil *jsem vidět ho zblízka, anebo se aspoň zeptat, kdo to je a co znamená.*
Strasznie ***chciałem*** *zobaczyć go z bliska albo przynajmniej się spytać,*
kto to jest i co to znaczy.

(13) *toužit* Inf → *pragnąć* Inf (44 výskytů — 30 %)
Dobře děláš, řekl náhle v obluzení, ***touže*** *ji zlíbat a cítě strach.*
Dobrze robisz — rzekł jak urzeczony, ***pragnąc*** *ucałować ją*
i czując jednocześnie strach.

(14) *toužit* Inf → *marzyć o* Oa (4 výskyty — 5 %)
Netoužila *o něm dlouze rozprávět.*
Nie marzyła *o długiej rozmowie na ten temat.*

2.5 *TOUŽIT (PO)* S

(15) *toužit* S → *pragnąć* S (13 výskytů — 57 %)
Celý život ***toužil****, aby milovaná žena byla s to tlouci kvůli němu hlavou o zeď,*
křičet zoufalstvím anebo skákat radostí po pokoji.
Przez całe życie ***pragnął****, żeby ukochana kobieta gotowa była bić dla niego głową*
w mur, wyć z rozpaczy i skakać z radości po mieszkaniu.

(16) *toužit* S → *chcieć* Inf (4 výskyty — 17 %)
Toužila*, aby s ní sdílelo její samotu alespoň nějaké zvířátko.*
Chciała *dzielić z kimś swą samotność, choćby z jakimś zwierzątkiem.*

(17) *toužit po* S → *chcieć* S (2 výskyty — 9 %)
Obama ***netouží*** *po tom, aby se problémy evropského dluhu rozšířily do Ameriky.*
Prezydent Obama ***nie chce*** *przecież, żeby europejski problem długu państwowego*
przeniósł się do Ameryki.

3. SHRNUTÍ RUČNÍ ANALÝZY

Výsledky ruční analýzy shrnuje tabulka 2, pro srovnání i s údaji o českých textech přeložených z polštiny.[18] Typy objektu v češtině jsou ve druhém řádku tabulky, odděleně pro české a polské originály. První dva sloupce obsahují ekvivalenty, tj. lemma a typ objektu v polštině. Např. sloveso *toužit*, v českých originálech ve spojení s abstraktním objektem (*toužit po* Oa), má v polských překladech 20 dokladů ekvivalentu *marzyć o* Oa. Stejné spojení (*marzyć o* Oa) v polských originálech přeložených do češtiny jako *toužit po* Oa má 5 výskytů. I když je celkový počet výskytů *toužit* v přeložených textech výrazně nižší (145 oproti 219), tendence volby ekvivalentů jsou podobné jako u opačného směru překladu, zvlášť u častěji zastoupených ekvivalentů, jako je třeba *pragnąć*.[19]

18 V 5. vydání InterCorpu je počet slov v českých a polských textech téměř stejný — asi 1,8 milionu.

19 Původní české texty vykazují vyšší frekvence lexému *toužit* než překlady do češtiny obecně, alespoň na základě dat z korpusu InterCorp (Kaczmarska — Rosen, 2013). Zatímco u polských textů je poměr frekvence *toužit* v originálech k překladům 1,67 a u německých 1,87, u ruských textů je to 2,37, a ve španělštině dokonce 3,13. Částečně to lze vysvět-

toužit (po)		**České originály**						**Polské originály**						
Polský ekvivalent	Valence	**Σ**	Oa	Oh	Or	Inf	S	**Σ**	Oa	Oh	Or	Inf	S	–
chcieć	Oa	0						3	2				1	
chcieć	Inf	23	1		1	20	1	9	1			6	2	
chcieć	S	2					2	0						
dążyć do	Oa	2	2					0						
łaknąć	Or	0						2			2			
marzyć o	Oa	24	20			4		6	5			1		
marzyć o	Oh	3	1	2				0						
marzyć o	Or	5			5			1			1			
marzyć	S	3			1		2	0						
mieć ochotę	Inf	2		1		1		1				1		
pożądać	Oa	5	5					3	3					
pożądać	Oh	5		5				2		2				
pragnąć	Oa	33	29	1		3		26	25				1	
pragnąć	Oh	12		12				2		2				
pragnąć	Or	3			3			1			1			
pragnąć	Inf	53	4	1	1	44	3	24	1			23		
pragnąć	S	15	1			1	13	3					3	
pragnąć		1		1				4						4
pragnienie	Oa	0						3				3		
(s)próbować	Inf	0						2				2		
spragniony	Oa	0						3	3					
tęsknić do	Oa	12	11	1				2	2					
tęsknić do	Oh	5		5				1		1				
tęsknić za	Oa	7	7					1	1					
tęsknić za	Oh	2		2				0						
tęsknić za	Or	2			2			0						
złakniony	Oa	0						2	2					
JINÉ		27	9	7	2	7	2	35	10	2	2	20		1
Σ		246	90	38	15	80	23	136	55	7	6	56	7	5

TABULKA 2. Polské ekvivalenty českého slovesa *toužit* — shrnutí

Ruční analýza částečně potvrdila hypotézu, že valence ovlivňuje volbu polského ekvivalentu. U slovesa *toužit* však lze ekvivalent spolehlivě určit jen ve spojení s infinitivem (v 81 % případů), tedy pro vzorec *toužit* Inf → *pragnąć / chcieć / mieć ochotę* Inf.[20] V ostatních typech výsledky nebyly průkazné.[21] U abstraktních objektů je velký rozptyl ekvi-

lit skladbou textů (tj. značným zastoupením děl Milana Kundery mezi českými originály), ale prostor zůstává i pro zdůvodnění založené na problematičnosti volby tohoto lexému při překladu do češtiny.

20 Výrazy *pragnąć / chcieć / mieć ochotę* považujeme za synonymní. Hlavní rozdíl je v intenzitě pocitu.

21 Nezkrácené výsledky viz Kaczmarska — Rosen (2013).

valentů zvlášť markantní a ukazuje na potřebu hlubší analýzy objektů. Jako test jsme zvolili dva abstraktní objekty — *velká láska* (*wielka miłość*) a *exotická cesta* (*egzotyczna podróż*). Se slovesem *toužit* se oba spojují snadno: *toužit po velké lásce / exotické cestě*. Znatelně vybíravější jsou v kombinaci se třemi nejčastějšími polskými ekvivalenty:

(18) *Marzyć o wielkiej miłości / egzotycznej podróży*
Tęsknić za wielką miłością / egzotyczną podróżą (?)
Tęsknić do wielkiej miłości / egzotycznej podróży (?)
Pragnąć wielkiej miłości / egzotycznej podróży (?)

Oba objekty jsou stejně přijatelné se slovesem *marzyć*, ale ani jeden s *tęsknić*.[22] Sloveso *pragnąć* je také méně přijatelné ve spojení s „exotickou cestou". Nejednoznačné výsledky ruční analýzy si žádají podrobnější zkoumání, která by výsledky ověřila, případně odhalila další faktory ovlivňující volbu ekvivalentů.

4. AUTOMATICKÁ EXCERPCE EKVIVALENTŮ

Při ruční analýze v části 2 jsme se omezili na několik lexémů a původní české texty. K ověření výsledků jsme provedli automatickou extrakci dvojic lexémů[23] ze všech česko-polských textů v jádru 6. vydání *InterCorpu*, a to bez ohledu na směr překladu (asi 12 mil. slov na české i polské straně).[24] Výsledkem bylo 8,7 milionu dvojic lemmat, z nichž 0,5 milionu dvojic bylo jedinečných, celkem se 121 tisíci českými a 98 tisíci polskými lemmaty.[25] Každému zarovnanému českému lexému ze vstupních textů tedy byla přiřazena množina ekvivalentů s frekvencí výskytu u každého z nich. Česko-polský glosář lze snadno proměnit na polsko-český setříděním dvojic podle polského sloupce. V příkladu (19) uvádíme seznam 16 nejčastějších ekvivalentů slovesa *toužit*, setříděný podle frekvence (v závorkách).

22 Ve spojení s *tęsknić* je spojení přijatelné jen v případě, že „velkou láskou" je lidská bytost, a nikoli abstraktní pojem.

23 Podrobnější údaje o metodě zarovnání po slovech viz Och — Ney (2003), byly použity jen věty zarovnané 1 : 1 a nejpřísnější nastavení, které vybírá jen spolehlivěji určené dvojice ekvivalentů. Alternativním nástrojem může být např. http://code.google.com/p/berkeleyaligner/. Polské texty byly lemmatizovány nástroji *Morfeusz* (http://sgjp.pl/morfeusz/) a *TakIPI* (http://nlp.pwr.wroc.pl/takipi/, viz Piasecki, 2007), české nástrojem *Morče* (http://ufal.mff.cuni.cz/morce/index.php, viz Votrubec, 2006). Z jiných projektů překladových slovníků na základě paralelního korpusu lze uvést např. Skoumalová (2008) nebo Jirásek (2011). Přehled o možnostech extrakce dvoujazyčných slovníků z paralelních a srovnatelných korpusů obsahuje Sharoff et al. (2013). Česko-polský glosář už posloužil pro kontrastivní lexikální studii deminutiv, viz Rosen et al. (2014). Metoda samotná byla použita i na další dvojice jazyků z *InterCorpu*. Výsledky jsou dostupné na http://treq.korpus.cz.

24 Tj. včetně textů, jejichž originál je v jiném jazyce — větší objem textů v novějším vydání, navíc bez omezení na jazyk originálu, totiž zvyšuje spolehlivost zarovnání po slovech.

25 Rozdíl v počtu lemmat je způsoben zejména odlišnými zásadami lemmatizace v češtině a polštině.

(19) *pragnąć* (304), *chcieć* (107), *tęsknić* (82), *marzyć* (70), *pożądać* (40), *ochota* (24), *zapragnąć* (9), *pragnienie* (8), *tęsknota* (8), *zależeć* (8), *spragniony* (7), *życzyć* (6), *upragniony* (5), *chęć* (4), *szukać* (4), *zatęsknić* (4)

V tabulce 3 porovnáváme relativní frekvence některých častějších ekvivalentů slovesa *toužit*. Frekvence byly zjištěny jednak na základě ruční excerpce z dat omezených na české a polské originální texty (sloupec 1 a 2), jednak na základě automatické excerpce ze všech beletristických česko-polských textů korpusu *InterCorp*, včetně překladů z jiných jazyků (sloupec 3). Z údajů vyplývá, že rozdíly v zastoupení jednotlivých ekvivalentů ve výsledcích ruční metody použité na menších a přísněji vybraných datech a automatické na větších a méně restriktivně vymezených datech nejsou zásadní.

	ruční excerpce z polských překladů českých originálů	ruční excerpce z polských originálů	automatická excerpce z polských originálů i překladů
chcieć	10,50 %	12,73 %	15,46 %
marzyć	15,13 %	8,18 %	9,93 %
ochota	0,84 %	0,91 %	3,40 %
pożądać	4,20 %	4,55 %	5,96 %
pożądany	0,42 %	0,91 %	0,28 %
pragnąć	49,58 %	54,55 %	43,69 %
pragnienie	1,26 %	3,64 %	1,13 %
tęsknić	13,45 %	4,55 %	11,63 %
zapragnąć	1,26 %	0,91 %	1,28 %

TABULKA 3. Relativní frekvence polských ekvivalentů slovesa ***toužit*** při ruční a automatické excerpci

5. DISKUSE A PERSPEKTIVY

Metodu ruční analýzy a automatické extrakce, podobně jako všechny další metody zkoumání lexikálních ekvivalencí v paralelních textech, lze použít na libovolné relevantní lexémy, tedy nejen na obtížné případy typu *toužit*, kde se hypotéza o vlivu valence na volbu polského ekvivalentu potvrdila jen částečně.[26] Konkrétně u slovesa *toužit* je možné spolehlivě určit ekvivalent jen ve spojení s infinitivem. U jiných tříd objektů výsledky nebyly průkazné a ukázaly na potřebu hledání dalších faktorů, nejlépe opět spojením hlubší analýzy paralelních konkordancí na menším vzorku dat s jejím doplněním a ověřením na větších datech automatickou metodou. Hlavním důvodem je nepochybně fakt, že volba ekvivalentu u sloves citového vnímání při překladu z češtiny do polštiny patří mezi velmi náročné úkoly i pro překladatele. Rozdíly mezi jednotlivými významovými odstíny jsou často minuciózní nebo obtížně uchopitelné, a rodilí mluvčí tato slovesa dokonce někdy ani jako polysémní nevnímají. Proto

26 Podobně byla zkoumána také slovesa *mrzet*, *být líto* a *mít rád*, viz Kaczmarska (2015a, 2015b).

je namístě uvažovat i o jiných faktorech, které se modelují a parametrizují hůře než argumentová struktura: širší kontext, situace nebo styl. Tak například může hrát roli skutečnost, zda entita nebo událost vyjádřená daným argumentem už existuje nebo existovala v minulosti, zda s ní subjekt už někdy přišel do styku. V takovém případě se *toužit* přeloží spíše jako *tęsknić* (např. *toužím po moři → tęsknię za morzem*). V případě opačném, kdy se něco může stát nebo objevit v budoucnosti nebo nikdy, je namístě spíše *marzyć* (*toužím být zdravá → marzę, żebym była zdrowa*). Snáze zjistitelnou okolností může být rod mluvčího, např. *kochać* (milovat) je pravděpodobně vhodnější ekvivalent pro *mít rád* než *lubić* (líbit se, mít rád), pokud je mluvčím žena.

Automatické metody jsou perspektivní především tím, že se dají použít na mnohem větší objemy textů. Automaticky pořízený seznam excerpovaných ekvivalentů však kromě frekvence nenabízí žádné vodítko, který z nich je v daném kontextu nejvhodnější. K doplnění a ověření závěrů ruční analýzy je třeba sofistikovanější metoda, která bere v úvahu kontext. Takový úkol může zvládnout i standardní stochastický klasifikátor, jehož cílem je predikovat nejvhodnější polský lexikální ekvivalent na základě českého lexému a jeho kontextu (Kaczmarska et al., 2015). Kontext je přitom možné chápat různě: (i) jako lineární posloupnost několika lemmat vlevo a vpravo od zkoumaného lexému nebo (ii) v podobě syntakticky závislých větných členů, přesněji jako jejich funkčně a slovnědruhově identifikované hlavy spolu s lemmaty. Metoda současně poskytuje i údaje o tom, které údaje mají pro volbu ekvivalentu nejvyšší vypovídací hodnotu (*information gain*). I když ani výsledky automatických metod nelze v této fázi interpretovat jednoznačně, optimalizací parametrů, kvalitnějšími daty a detekcí dalších údajů z textu směřujeme k podchycení a evaluaci dalších potenciálních faktorů, a tím pádem i k co nejlepší predikci lexikálního ekvivalentu. Pokud by bylo možné dospět k signifikantním a zároveň algoritmicky identifikovatelným faktorům,[27] byly by výsledky využitelné např. pro strojový překlad (viz např. Bojar, 2012; Han et al., 2013).

LITERATURA

Baker, M. (1992): *In Other Words: A Coursebook on translation*. London — New York: Routledge.

Bojar, O. (2012): *Čeština a strojový překlad*, Studies in Computational and Theoretical Linguistics 11. Praha: ÚFAL MFF UK.

Catford, J. C. (1965): *A Linguistic Theory of Translation: An Essay In Applied Linguistics*. London: Oxford University Press.

Čermák, F. — Rosen, A. (2012): The case of InterCorp, a multilingual parallel corpus, *International Journal of Corpus Linguistics* 13, 3, s. 411–427.

Čermáková, A. (2009): *Valence českých substantiv*. Praha: Nakladatelství Lidové noviny.

Dąmbska-Prokop, U. (2000): *Mała encyklopedia przekładoznawstwa*, Częstochowa: EDUCATOR.

Daneš, F. — Hlavsa, Z. (1987): *Větné vzorce v češtině*. Praha: Academia.

Dębski, A. (1982): Semantyczna walencja czasownika w aspekcie konfrontatywnym, *Biuletyn Polskiego Towarzystwa Językoznawczego*, 39, s. 79–90.

27 Např. sémantické třídy lze aproximovat pomocí dostatečně reprezentativního tezauru. Jedním z možných kandidátů je WordNet (http://hdl.handle.net/11858/00-097C-11858/00-097C-0000-0001-487A-4, http://plwordnet.pwr.wroc.pl/wordnet/).

Greń, Z. — Rytel-Kuc, D. (1991): Wykorzystanie przekładów literackich w pracy nad dwujęzycznym słownikiem walencyjnym. In: H. Běličová et al. (eds.), *Problemy teoretyczno-metodologiczne badań konfrontatywnych języków słowiańskich*. Warszawa: Instytut Słowianoznawstwa PAN, s. 69–78.

Han, A. L. — Lu, Y. — Wong, D. F. — Chao, L. S. — He, L. –Junwen, X.(2013): Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*, ACL, s. 365–372.

Havránek, B. — Bělič, J. — Helcl, M. — Jedlička, A. (eds.) (2011): *Slovník spisovného jazyka českého*. Praha: ÚJČ.

Hejwowski, K. (2009): *Kognitywno-komunikacyjna teoria przekładu*, Warszawa: PWN.

Jirásek, K. (2011): Využití paralelního korpusu InterCorp k získávání ekvivalentů pro chorvatsko-český slovník. In F. Čermák (ed.), *Korpusová lingvistika Praha 2011: 1 — InterCorp*. Praha: NLN, s. 45–55.

Jelínek, T. (2014): Improvements to dependency parsing using automatic simplification of data. In: N. Calzolari et al. (eds.), *Proceedings of LREC'14*. Reykjavík: ELRA, s. 73–77.

Kaczmarska, E. (2001): Badanie struktury walencyjnej czeskich i polskich predykatów posiadających pozycję Experiencera. *Studia z Filologii Polskiej i Słowiańskiej* 37. Warszawa: Slawistyczny Ośrodek Wydawniczy, s. 177–187.

Kaczmarska, E. (2015a): W poszukiwaniu znaczenia czasowników wyrażających stany psychiczne. *Prace Filologiczne*. (V tisku).

Kaczmarska, E. (2015b): Czeskie czasowniki oznaczające stany psychiczne — sposoby ustalania polskich ekwiwalentów na podstawie korpusu równoległego InterCorp. *ZBLIŻENIA*, 13.–14. 11. 2013, Konin, Polsko.

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. Warszawa: Slawistyczny Ośrodek Wydawniczy, s. 103–121.

Kaczmarska, E. — Rosen, A. (2014a): Praktyczny przewodnik po korpusie równoległym InterCorp. In: M. Hebal-Jezierska (ed.), *Praktyczny przewodnik po korpusach języków słowiańskich*. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego, s. 207–231.

Kaczmarska, E. — Rosen, A. (2014b): Czego nie można wyrazić w języku polskim, czyli o leksykalnych w nim brakach. *Polonica,* 34, Instytut Języka Polskiego PAN, s. 53–66.

Kaczmarska, E. — Rosen, A. — Hana, J. — Hladká, B. (2015): Syntactico-semantic analysis of arguments as a method for establishing equivalents of Czech and Polish verbs expressing mental states. *Prace Filologiczne* (v tisku).

Koller, W. (1995): The concept of equivalence and the object of translation studies, *Target* 7, 2.

Levin, B. (1993): *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Lewandowska-Tomaszczyk, B. (1984): *Conceptual Analysis, Linguistic Meaning, and Verbal Interaction*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Lewandowska-Tomaszczyk, B. (2013): Komunikacja i konstruowanie znaczeń w przekładzie. *ZBLIŻENIA*, 13.–14. 11. 2013, Konin, Polsko.

Lopatková, M. — Kettnerová, V. — Bejček, E. — Skwarska, K. — Žabokrtský, Z. (2014): *VALLEX 2.6.3 — Valency Lexicon of Czech Verbs*. Praha: ÚFAL MFF UK. http://ufal.mff.cuni.cz/legacy/vallex/2.6.3/doc/home.html.

Lotko, E. (1992): *Zrádná slova v polštině a češtině (Lexikologický pohled a slovník)*. Olomouc: Votobia.

Nida, E. A. (1995): Dynamic Equivalence In Translating. In *An Encyclopaedia of Translation. Chinese-English / English-Chinese*, Hong Kong: Chinese University of Hong Kong.

Nivre, J. — Hall, J. (2005): MaltParser: A language-independent system for

data- driven dependency parsing. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, s. 13–95.

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

Oliva, K. (1994). *Polsko-český slovník*. Praha: Academia.

Piasecki, M. (2007): Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly* 11, 1–2, s. 151–167.

Rosen, A. — Kaczmarska, E. — Škodová, S. (2014): Zdrobnienia jako element kultury i pułapka glottodydaktyczna — czeskie i polskie deminutiva w ujęciu konfrontatywnym na podstawie badań korpusowych. In: E. Kaczmarska — A. Zieniewicz (eds.), *Glottodydaktyka wobec Wielokulturowości*. Warszawa, s. 51–63.

Rytel, D. (1989): Wybrane problemy opisu walencyjnego języka. *Studia z Filologii Polskiej i Słowiańskiej* 26. Warszawa: Slawistyczny Ośrodek Wydawniczy, s. 237–247.

Rytel-Kuc, D. (ed.). (1991): *Walencja czasownika a problemy leksykografii dwujęzycznej*. Wrocław: Zakład Narodowy im. Ossolińskich.

Sharoff, S. — Rapp, R. — Zweigenbaum, P. — Fung, P. (eds.) (2013): *Building and Using Comparable Corpora*. Springer.

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

Skoumalová, H. (2008): Extracting dictionaries from parallel corpora. In *Proceedings of The Third Baltic Conference on Human Language Technologies*. Kaunas: Vytautas Magnus University, s. 297–301.

Urbańczyk-Adach, N. (2011): *Wariantywność walencji czeskiego czasownika*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Votrubec, J. (2006): Morphological tagging based on averaged perceptron. In *WDS'06 Proceedings of Contributed Papers*, Praha: Matfyzpress, Univerzita Karlova v Praze, s. 191–195.

Elżbieta Kaczmarska | Instytut Slawistyki Zachodniej i Południowej, Uniwersytet Warszawski | Krakowskie Przedmieście 26/28, 00–927 Warszawa
e.h.kaczmarska@uw.edu.pl

Alexandr Rosen | Ústav teoretické a komputační lingvistiky, FFUK | Celetná 13, 110 00 Praha 1
alexandr.rosen@ff.cuni.cz