Mnohojazyčný korpus InterCorp: Možnosti studia

František Čermák a Jan Kocek (eds)

NAKLADATELSTVÍ LIDOVÉ NOVINY

Ústav Českého národního korpusu

Publikace vychází v rámci výzkumného záměru MSM0021620823 *Český národní korpus a korpusy dalších jazyků* řešeného na FF UK.

Recenzenti:
prof. PhDr. Eva Hajičová, DrSc.
prof. PhDr. Alena Macurová, CSc.

ISBN 978-80-7422-058-6

(A)symetrie valenčních vlastností českých a anglických sloves pohybu

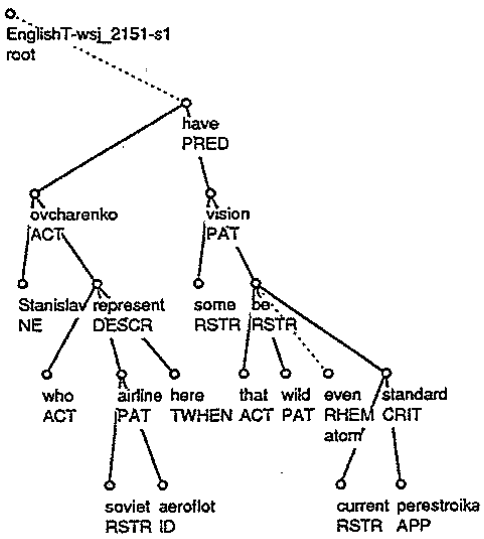
Jana Šindlerová
Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova v Praze

1. Úvod

Novější teorie slovesné valence, jako je konstrukční gramatika (viz např. (Goldberg 1995)), přicházejí s myšlenkou, že valenční vlastnosti slovesa nejsou inherentně vlastní slovesu jako lexikální jednotce, ale jsou v zásadě záležitostí kontextu, ve kterém je sloveso použito. Konkrétní naplnění valenčních rámců a realizace sémantických rolí v textu jsou podle nich dány působením generalizovaných, abstraktních syntaktických struktur, odrážejících základní dějové situace v realitě. Tyto abstraktní syntaktické struktury mohou být jak jazykově univerzální, tak jazykově specifické a přispívat tak k podobné, či různé syntaktické stavbě ekvivalentních úseků textu. Máme-li se tedy zabývat valenčními vlastnostmi sloves, je nezbytné studovat je na autentickém jazykovém materiálu, ideálně pak na anotovaném korpusu. Chceme-li valenční vlastnosti sloves zkoumat v perspektivě mezijazykových vztahů, potřebujeme korpusy paralelní.

V našem příspěvku se budeme zabývat srovnáním valenčních vlastností českých a anglických sloves v textech ze dvou formálně odlišných zdrojů: korpusu Prague Czech-English Dependency Treebank (dále jen **PCEDT**) a korpusu **Intercorp.** PCEDT je dvojjazyčný treebank anotovaný na více jazykových rovinách (viz obrázek 1). Obsahuje 49208 anglických vět z ekonomického deníku Wall Street Journal a jejich překlady do češtiny. Syntaktická anotace PCEDT, včetně informace o valenci, je zachycena hlavně na tzv. tektogramatické rovině, tj. rovině hloubkových syntakticko-sémantických vztahů.

Součástí korpusu PCEDT jsou dva elektronické valenční slovníky, PDT-VALLEX a Engvallex. **PDT-VALLEX** je elektronický valenční slovník českých sloves vyvinutý v rámci anotačních prací na Pražském závislostním korpusu (PDT). Pro každé slovesné lemma tento slovník uvádí seznam možných valenčních rámců, přičemž v zásadě platí, že každý rámec odpovídá samostatnému slovesnému významu. Valenční rámec je zapsán jako posloupnost valenčních doplnění, vyjádřených konvenčními značkami, tzv. funktory. Za valenční doplnění relevantní pro zápis do rámce jsou považována obligatorní i fakultativní vnitřní doplnění a obligatorní volná doplnění (k definici pojmů vnitřní/volná doplnění a obligatorní/fakultativní doplnění viz (Panevová 1980)). U jednotlivých valenčních doplnění je uveden výčet typických forem, v nichž se toto doplnění v korpusových datech vyskytuje. Pro každý rámec jsou k dispozici příkladové věty, obvykle obsahující informace o dalších, fakultativních volných doplněních, s nimiž se dané sloveso v tomto významu obvykle vyskytuje.



Stanislav Ovcharenko, who *T*-1 represents the Soviet airline Aeroflot here, has some visions that *T*-2 are wild even by the current standards of perestroika.

Obrázek 1: Zachycení tektogramatické syntaktické struktury v anglické části korpusu PCEDT

Engvallex byl vytvořen konverzí valenčního slovníku PropBank (Kingsbury et al. 2002) do teoretického formátu Funkčního generativního popisu (Sgall 1986). V rámci konverze proběhlo přejmenování participantů konvenčními značkami tektogramatické roviny korpusu PDT, vymazání fakultativních volných doplnění, slučování rámců s identickým slovesným významem, doplnění rámců o další participanty podle pravidel FGP a přidání dalších potřebných rámců. Dále byla provedena celková kontrola rámců Engvallexu a v případě nejasností byla struktura rámce přiblížena struktuře odpovídajícího rámce ve slovníku PDT-VALLEX.

Oba výše zmíněné valenční slovníky mají podobnou strukturu a využívají jednotného uživatelsko-anotátorského prostředí. Datová struktura obou slovníků je reprezentována ve formátu XML. V současné době se provádí propojení těchto dvou slovníků do dvojjazyčného elektronického překladového slovníku (Šindlerová 2009). Propojování slovníků probíhá jako propojování a) jednotlivých valenčních rámců a b) jednotlivých participantů v rámcích. Pro každý valenční rámec uvedený v Engvallexu je vytvořen seznam odpovídajících rámců PDT-VALLEXU a v rámci každé položky tohoto seznamu je zapsáno konkrétní propojení jednotlivých participantů.

Korpus **Intercorp** je multilingvální korpus literárních textů a jejich překladů. Je mnohojazyčný, ale neobsahuje syntaktickou anotaci, tudíž primárně ani explicitní značení slovesné valence. Jeho přínos pro zkoumání valenčních vlastností tkví

v jeho velikosti, univerzálnosti a v širokém tematickém záběru obsažených textů. Domníváme se, že právě vzájemná formální i obsahová různorodost obou zdrojových korpusů významně přispívá k ucelenosti pohledu na dané téma.

2. PCEDT versus Intercorp

V tabulce 1 shrnujeme nejdůležitější aspekty srovnání korpusů PCEDT a Intercorp.

	PCEDT	INTERCORP_CZ/EN
velikost	49 208 vět = cca 1,2 mil. slov	cca 2 376 075/ 2 834 404 slov (v době vzniku příspěvku)
druh textů	publicistické, ekonomická specializace (Wall Street Journal)	beletristické
směr překladu	Eng → Cz	Eng ↔ Cz
charakter anotace	morfologická, syntaktická analytická (automaticky), syntaktická tektogramatická	morfologická
druh vyhledávání	na základě morfologických značek, na základě syntaktických vztahů a struktur	na základě morfologických značek

Tabulka 1: Srovnání vlastností korpusů PCEDT a Intercorp

2.1. Velikost korpusu

Tektogramatická anotace korpusu PCEDT stále probíhá. V době konání konference projektu Intercorp bylo již dokončeno 63,80 % procent anotace české části a 40,31 % anglické části PCEDT. Průnik paralelní anotace v obou částech tvořil 11547 vět, tj. 269 035 českých a 290 919 anglických slov. Korpus Intercorp obsahoval ve stejné době cca 2,5 mil. slov na každé straně, tj. přibližně desetkrát větší množství dat.

Přestože je rozdíl současného objemu Intercorpu a PCEDT několikanásobný, nezaznamenali jsme výraznější rozdíl v počtu vzájemně různých překladových ekvivalentů v obou korpusech. Pro ilustraci uvádíme příklad slovesa *think*, které se v hotových datech korpusu PCEDT vyskytlo ve 108 případech a bylo přeloženo 19 různými českými slovesy. V korpusu Intercorp se sloveso *think* objevilo ve 286 případech a přeloženo bylo 27 různými slovesy či verbonominálními spojeními. Rozdíl mezi počty ekvivalentů podle nás spočívá spíše ve větší tematické i stylové variabilitě textů korpusu Intercorp než v objemu jeho dat.

2.2. Druh textů

PCEDT obsahuje texty specificky tematicky zaměřené. Jedná se převážně o žurnalistické texty ekonomického zaměření, popřípadě o krátké zprávy ze světového

dění. To významně ovlivňuje frekvenci slovesných významů, které se v korpusu objevují. Velký výskyt můžeme očekávat u sloves mluvení, sloves obchodní transakce a sloves pohybu. U ekonomických textů předpokládáme vyšší frekvenci sloves centrální slovní zásoby, popřípadě terminologických sloves.

Intercorp obsahuje texty převážně beletristické, slovní zásoba v nich tedy není nijak významně omezena. Na rozdíl od PCEDT, kde se vyskytuje i značný podíl metaforičnosti, ale zde nepůjde jen o metaforičnost žurnalisticky konvenční. Očekáváme větší překladovou variabilitu.

2.3. Směr překladu

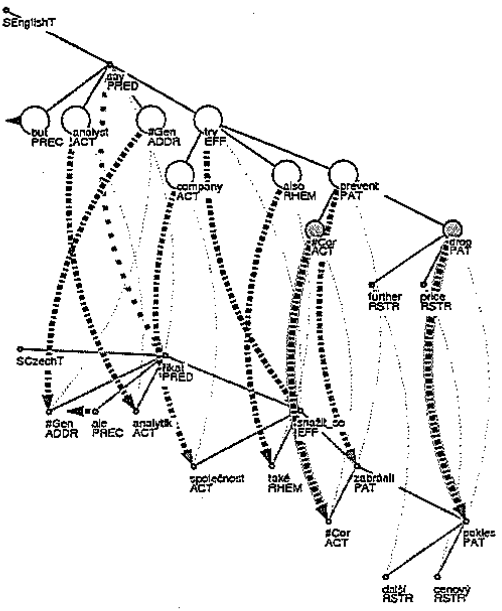
PCEDT obsahuje jednosměrně přeložené věty: anglické texty byly přeloženy do češtiny. Překlady byly pořízeny přímo pro účely projektu a byly několikrát revidovány. Prioritou překladu byla věrnost originálu, a to i po stránce formální. Naopak korpus Intercorp shromáždil texty přeložené nezávisle, neúčelově a vzhledem k jejich charakteru víceméně volně. Na rozdíl od korpusu PCEDT zde tedy dochází často k situaci, kdy sloveso zdrojové věty v cílové větě nemá (ani nominální) ekvivalent. Intercorp obsahuje také texty přeložené z češtiny do angličtiny. To představuje zajímavý prvek pro zpětnou vazbu při hodnocení ekvivalentnosti překladů. Při porovnání výsledků s korpusem PCEDT by však tyto texty neměly být brány v potaz.

2.4. Charakter anotace

PCEDT obsahuje stejné roviny anotace jako PDT 2.0, ovšem důraz je kladen na anotaci tektogramatické roviny, tj. roviny hloubkově syntaktických vztahů. Na této rovině je také každému slovesu přiřazena informace o odpovídajícím valenčním rámci. V samotné anotaci konkrétních vět pak pod každým slovesem musí být přítomny všechny obligatorní členy rámce. Podrobná syntaktická anotace umožňuje snadno identifikovat členy závislé na slovesu a mezi nimi pak i členy valenční.

Pro PCEDT je v projektu multilingválního valenčního slovníku k dispozici zobrazení automatického zarovnání uzlů (node alignment), včetně valenčních pozic, mezi odpovídajícími větami (viz obrázek 2). To uživateli významně usnadňuje práci na jakékoli srovnávací studii.

Intercorp syntaktickou anotaci neobsahuje, je anotován pouze morfologicky. Také není možné zobrazit zarovnání uzlů, tudíž je srovnávací práce ztížena nutností ručně procházet každou větu a určovat odpovídající uzly. Vzhledem k tomu, že již Pražský závislostní korpus byl připravován na materiálu ČNK, automatická transformace Intercorpu do (alespoň částečně relevantních) stromových struktur a aplikace zarovnávací procedury by s odpovídajícím softwarem pravděpodobně nebyla náročným projektem.



Obrázek 2: Automatické zarovnání uzlů v nástroji pro tvorbu multilingválního valenčního slovníku

2.5. Druh vyhledávání

PCEDT je díky svému zaměření na syntaktickou anotaci vhodný pro vyhledávání valenčních vazeb. Kromě syntaktické anotace lze ovšem využít i vyhledávání podle morfologických vlastností jednotlivých uzlů. Prohledávání korpusu je v současné době možné pomocí různých nástrojů (např. (Pajas et al. 2009)). Tyto nástroje jsme však zatím nevyužili, protože v projektu multilingválního valenčního slovníku již přímárně pracujeme se zarovnanými soubory setříděnými podle slovesných lemmat a frekvence výskytu v korpusu. V první fázi výzkumu se zaměřujeme na sledování rozdílů ve valenci slovesných ekvivalentů, nástroje na dotazování budou vhodné zejména v dalších fázích, kdy se již nebudeme zabývat přímo slovesnými lemmaty, ale jednotlivými druhy slovesných participantů, tj. kdy přesuneme hledisko výzkumu „o patro níže".

Prohledávání Intercorpu probíhá na speciálním webovém rozhraní. K vyhledávání lze využít lemmatu, slovní formy či morfologického tagu, možnosti kombinace dotazů však zatím nejsou tak rozsáhlé jako v jiných uživatelských prostředích pro zpracování korpusů ČNK. Jak jsme již zmínili, korpus (zřejmě i vzhledem k volnosti překladů) neposkytuje informaci o zarovnání uzlů. Webové rozhraní také zatím bohužel neumožňuje jednoduché ukládání výsledků v přijatelném formátu.

2.6. Shrnutí

Srovnání obou korpusů ukazuje jejich přednosti i slabiny pro daný úkol. Korpus PCEDT nabízí lepší prohledávání, nástroje, syntakticky věrnější překlad a vlastní zdroje valenčních charakteristik sloves. Oproti tomu Intercorp má významnou výhodu v různorodosti a tematické neomezenosti textů.

3. Směrová doplnění

V tomto příspěvku pracujeme se specifickou skupinou sloves, vyjadřujících pohyb a obsahujících ve valenčním rámci jedno nebo více tzv. směrových doplnění. Na tuto skupinu sloves se zaměřujeme proto, že jejich valenční rámce bývají bohaté a různorodé, ale zároveň se jedná o homogenní sémantickou třídu.

Směrovými doplněními myslíme doplnění ve FGP označovaná funktory DIR1, DIR2 a DIR3. Tato doplnění vyjadřují v širokém smyslu směrové relace v prostoru, tedy odpovídají na otázky *odkud?*, *kudy?* a *kam?*. V zahraniční literatuře jsou tato doplnění někdy označována jako tzv. vektorové argumenty a jejich sémantické pozice jsou označeny termíny *zdroj (source)*, *trasa (path)* a *cíl (goal)* pohybu. Tato doplnění mají zvláštní, přechodovou pozici v systému aktantů a volných doplnění, bývají klasifikovány jako tzv. argumenty druhého stupně (*secondary complements*, *second level arguments* atd.) (Hedberg et al. 2002). Směrová doplnění jsou prototypicky vyjadřována předložkovou skupinou, což je přibližuje adjunktům, tj. volným doplněním. V teorii FGP jsou na základě různých kritérií stále klasifikována jako volná doplnění.

4. Asymetrie valenční struktury překladových ekvivalentů

4.1. Pozorování

Při předběžném zkoumání valenčního chování překladových ekvivalentů sloves pohybu v korpusech Intercorp a PCEDT jsme narazili na tři typy asymetrií.

První typ asymetrií souvisí s rozdílnou povahou obou jazyků. Neflektivní, analytická angličtina nevyužívá obvykle ke specifikaci a derivaci slovesných významů prefixaci, ale spíše zachovává jednu lexikální jednotku, ke které připojuje v textu širokou škálu předložek a částic. Množství odpovídajících významů tak vzniká vlastně na úrovni syntaxe. Tyto významy se posléze mohou osamostatňovat a předložky či částice se postupně lexikalizují jako součást lemmatu. Mnohdy je pak obtížné míru lexikalizace v rámci lexikografické práce posoudit a rozhodnout, zda je vhodné slovesný význam zařadit ještě pod základní lexikální jednotku, nebo je pro něj nutné vytvořit frázové lemma.

Slovesné prefixy v češtině, stejně jako lexikalizované částice u frázových sloves v angličtině, mají jednu podstatnou vlastnost: často vlastně sémanticky fixují

jeden z argumentů jako obligatorní. Vzhledem k hojné derivaci nových slovesných lemmat pro specifické významy pomocí prefixace je tak čeština vlastně částečně valenčně určující již na úrovni lemmatu, zatímco angličtina až na úrovni syntaktického zapojení slovesa v kontextu. Důsledkem je větší variantnost překladových ekvivalentů v češtině, jak na úrovni lemmat, tak na úrovni rámců, a rozdíly v obligatornosti. Očekáváme tedy obligatornost v češtině i na takových místech, kde se v angličtině vyskytuje fakultativnost.

Druhým typem asymetrií jsou případy, kdy směrová (ale i lokační nebo měrová) valenční doplnění v češtině jsou v angličtině nahrazena patientovým argumentem. Je ovšem diskutabilní, do jaké míry je patientová klasifikace takových doplnění oprávněná. Hodnocení argumentů se zde opírá zejména o fakt, že jsou v angličtině vyjádřeny prototypicky přímým pádem, akuzativem. Na druhé straně hodnocení některých předložkových valenčních členů coby směrových (*uhodit do stolu*) se mnohdy opírá také spíše o jejich formu, než o sémantiku. Máme však za to, že zvolená forma v jazyce odráží pozměněné chápání těchto doplnění samotnými mluvčími daného jazyka.

Posledním typem asymetrie, který je evidentní již při zběžném pohledu na daný jazykový materiál, je formální podspecifikace distinkce DIR/LOC, tj. distinkce směru vs. lokace.

V následujícím oddílu budeme výše uvednené typy asymetrií dokládat pouze příklady z korpusu Intercorp, v korpusu PCEDT se však vyskytují v hojném počtu také.

4.2. Případová studie

4.2.1. Crawl

Crawl je typické sloveso vyjadřující způsob pohybu (manner of motion verb). U těchto sloves jsou typicky všechny tři vektorové argumenty fakultativní a často nejsou povrchově vyjádřeny vůbec.

(1) *"Nemohl lézt po čtyřech," odpověděla Rút.*
(1') *,He couldn't< crawl>,' Ruth said.*

Na slovese crawl bychom rádi ukázali typ asymetrie označený v tomto článku jako typ první. Při hledání výskytů slovesa crawl v korpusech nacházíme velké množství kombinací tohoto slovesa s předložkami: *crawl along, crawl forward, crawl in, crawl up, crawl out of, crawl into, crawl off, crawl through*, apod. Běžný slovník, jakým je např. Macmillan Dictionary (v online verzi), uvádí toto sloveso v podstatě bez frázových variant (jediná uvedená varianta *crawl with* by měla být hodnocena jako manifestace syntaktické alternace, nikoli zvláštní lexém). Každá z těchto variant je však přeložena jiným, většinou předponovým slovesem českým: *vyprostit se, zalézt, rozlézt se, protáhnout se*, apod.

(2) *A pained look <crawled> across one of Zaphod's faces and on to the other one.*
(2') *Po jedné ze Zafodových tváří přeběhl bolestný výraz a přešplhal na druhou.*
(3) *The window above her bathtub was too small for even a child to <crawl> through.*
(3') *Okno nad vanou bylo tak malé, že by se tamtudy neprotáhlo ani dítě.*

V obou výše uvedených případech se u českého slovesa ve valenčním rámci objeví obligatorní DIR2, zatímco v anglickém slovníku zůstane pouze fakultativním doplněním. Při propojování valenčních rámců tak potenciálně ztrácíme informaci o jinak pravidelném zarovnání uzlů v korpusu, pro obligatorní DIR2 na jedné straně nemáme ve slovníku dosažitelný ekvivalent na straně druhé, přestože v paralelních větách tento ekvivalent existuje. Pro lexikografa toto představuje otázku, zda informaci o zarovnání valenčního uzlu s uzlem nevalenčním do propojování rámců zahrnout (protože v praxi je užitečná), nebo ji pominout (protože pro teorii valence v rámci definice není relevantní).

U slovesa crawl se projevila v korpusu Intercorp i asymetrie třetího typu.

(4) *He didn't notice anything but the caterpillar bulldozers <crawling> over the rubble that had been his home.*
(4') *Nevšímal si ničeho kromě buldozerů plazících se po zbořeništi, které kdysi bylo jeho domovem.*

Namísto DIR2 v anglické větě byl v české větě použito lokativní určení (které je méně specifické a často zahrnuje i možné interpretace trajektorní). Asymetrie tohoto typu jsou zajímavé, protože se vyskytují pravidelně u sloves podobného typu a nastolují otázku, nakolik je valenční rozdíl skutečně zakotven v jazyce samém, či nakolik jde o osobní interpretaci autora překladu.

4.2.2. Descend

Sloveso *descend* představuje v některých výskytech také asymetrii prvního typu, ale v opačném směru překladu.

(5) *S pocitem, že už se vzpamatovala, se Enid vydala dolů na palubu B.*
(5') *Feeling restored, she<descended> to the „B" Deck.*

V sémantice slovesa descend je už inherentně obsažen obligatorní DIR1 (i DIR3), na rozdíl od rámce pro sloveso vydat, kde je podle PDT-VALLEXU obligatorní pouze DIR3.

Při sledování základní překladové dvojice *descend – sestoupit* je jasně vidět asymetrie druhého typu.

(6) *Vyšel z temné komory a pomalu sestoupil po mokrých schodech.*
(6') *He left the darkroom and <descended> the rain-slick stairway slowly.*

DIR2 je v češtině předložkový nebo vyjadřován nepřímým pádem. Na stejném místě ovšem v angličtině nalézáme řadu výskytů s pádem přímým, ukazujícím na patientovou interpretaci. Zdá se, že u řady sloves v angličtině může být interpretací DIR2 „zasaženost koridoru průchodem".

Tato vlastnost se produktivně šíří i k dalším slovesům s přeneseným významem pohybu.

(7) *Za mnou se pár lidí hrne po několika posledních schodech, aby to stihli, než se dveře zavřou.*
(7') *Behind me, other people <hustle> the last few steps to catch the door before it swings shut.*

To odpovídá předpokladům zmíněným ve studiích konstrukční gramatiky, že existují syntaktické konstrukce, které „cestují" mezi slovesy podobného významu, přenášejí se na základě analogie a nejsou tak vlastně přímo součástí lexikonu.

4.2.3. Travel

Sloveso travel v Engvallexu má pouze jediný, intranzitivní rámec. Kromě toho Macmillan Dictionary uvádí ještě tranzitivní variantu travel the world. V korpusu pak nalézáme příklady typu:

(8) *When she had< travelled> a hundred metres, Jimmy stopped pedalling.*
(8') *Když urazila sto metrů, Jimmy přestal šlapat.*

Ve formě přímého předmětu, který je typický pro patientové argumenty se zde vyskytuje doplnění vyjadřující míru vzdálenosti., které je v PDT v jiných formách obvykle značeno EXT. Totéž platí pro vazby typu *travel the world*, které jsou do češtiny obvykle překládány lokativní konstrukcí *cestovat po světě*.

Tato transformace neaktantového valenčního doplnění do přímého předmětu je pro angličtinu typická a pravděpodobně odráží změnu vnímání argumentu samotnými mluvčími. V češtině se častěji využívá předložkových vazeb, které jsou interpetovány jako volná doplnění. Přestože v češtině tato transformace není častá, u některých sloves se objevuje. Otázkou je, zda má vyjádření argumentu formou přímého předmětu skutečně i sémantický dosah, a pokud ne, tj. pokud bychom měli i takovéto akuzativní argumenty chápat stále jako volná doplnění, zda je nutné přehodnotit dosavadní pojímání tohoto jevu ve FGP.

Obligatornost jednotlivých argumentů je ve FGP určována pomocí tzv. *dialogového testu* (viz Panevová, 1980). Jeho princip spočívá v tom, že v dialogu je nemožné při otázce na obligatorní argument odpovědět: „Nevím." Dialogový test ovšem nefunguje stoprocentně. Např. angličtina, jak už jsme zmínili, má slovesné lemma často podspecifikované co do informace o trazitivnosti použití (tj. mnohdy lze sloveso v tomtéž (hrubém) významu použít v různých stupních tranzitivity). Zatímco dialogový test je nejúspěšnější s malým nebo téměř žádným kontextem, v angličtině je někdy kontext nezbytný pro úplné chápání slovesné valence.

(9) *In the spring through fall, one< travels> through the woods as if submerged in a sea of green [...]*
(9') *Od jara do podzimu pak člověk putuje lesy, jako by plul ponořený v moři zeleně [...]*

Travel jakožto sloveso pohybu umožňuje potenciální naplnění tří vektorových argumentů.

(10) *Football rowdies will travel from Austria to England through Czech Republic. [Coogle, upraveno]*

V příkladu (9) je však situace jiná. Zde je sémanticky relevantní pouze DIR2.

(9'') *In the spring through fall, one travels [*from New York] through the woods [*to Alaska] as if submerged in a sea of green [...]*

V tomto případě se nejedná o popis cesty se zdrojovou a cílovou lokací, ale o popis „bezcílného bloumání", podobně jako v jiných případech funguje LOC. Zde je však použita forma prototypická pro DIR2 a netypická pro LOC. Pro lexikografa je zde důležité rozhodnout, zda se skutečně jedná o rozšíření sémantiky DIR2, či o novou formu vyjádření LOC.

Prolínání směru a lokace se objevuje i v kontextově závislé interpretaci předložky *in*.

(11) *„We've< traveled> in Bulgaria," Alfred said.*
(11') *"Jednou jsme jeli do Bulharska," začal vykládat Alfred.*

V takových případech při nedostatku specifického kontextu vzájemně interferuje interpretace lokativní a cílová. Podobně jako v příkladě (9), i zde se jedná o dvojí chápaní slovesa *travel*, tj. o dvojí význam, zřetelný pouze z kontextu. První z těchto významů odpovídá situaci, kdy sloveso *travel* vyjadřuje obecné „vykonávání cesty", obvykle „pro potěšení"; v takových případech se v okolí tohoto slovesa objevuje fakultativní símanticky lokativní vyjádření.. Druhý význam odpovídá situaci, kdy subjekt „koná cestu s účelem", tato cesta má jasně daný start a cíl vyjádřený vektorovými argumenty: *He'd traveled with him to Florida to bury his mother.* Domníváme se, že v prvním případě je relevantní považovat sloveso *travel* za intranzitivní, zatímco v případě druhém je alespoň jeden z vektorových argumentů sémanticky obligatorní.

5. Závěr

České a anglické valenční struktury vypadají většinou velmi podobně, ale asymetrie mezi nimi existují. Tyto asymetrie představují výzvu pro jakýkoli teoretický rámec pro popis valence, který je vytvářen jako jazykově univerzální.

Při srovnávání valenčního chování sloves češtiny a angličtiny na materiálu anotovaných korpusů se ukazuje, že některé participanty mají možná jiný charakter, než bychom předpokládali na základě jednojazyčné zkušenosti. Také se ukazuje, že valence skutečně nemusí být jen záležitostí lexikonu. Pokud chceme vytvářet valenční slovníky, musíme do nich tento fakt jisté volnosti přenosu valenčních rámců zabudovat také. Poznání charakteru asymetrií překladových valenčních rámců je důležité i pro implementaci elektronických valenčních slovníků do systémů strojového překladu.

Pohled na valenční struktury v cizím jazyce nám může ozřejmit některé netypické chování valenčních struktur v jazyce vlastním. Je možné zkoumat míru zapojenosti jednotlivých valenčních struktur v jednotlivých jazycích, možnost jejich vzájemného ovlivňování, to, jak ovlivňují různé možnosti vyjádření význam překládaného o přeloženého. Náš výzkum je v této chvíli na samém začátku. Jedním z našich cílů je ověřit univerzálnost specifické valenční teorie, vytvořené při práci s českými daty, možnost jejího použití pro data jiných jazyků. Věříme, že korpusy, jako jsou PCEDT a Intercorp, nám tuto práci výrazně usnadní.

Tento článek vznikl za částečné podpory grantů GAUK 19008/2008 (Multilingvální zdroj valenčních vlastností sloves), MŠMT ČR LC536 (Centrum komputační lingvistiky) a ME838 (Reprezentace významu a automatické porozumění přirozenému jazyku (PIRE ČR)).

Literatura

Goldberg A.E., 1995, *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Hedberg N., R. DeArmond, 2002, On the Argument Structure of Primary Complements. In *Proceedings of the 2002 Annual Conference of the Canadian Linguistics Association*. Dostupné z WWW: <http://www.sfu.ca/~hedberg/CLA_2002.pdf>.

Kingsbury P., M. Palmer, 2002, From Treebank to Propbank. In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain.

Pajas P, J. Štěpánek, 2009, System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, eds. G.G. Lee, S. Schulte im Walde, Suntec, Singapore, 33-36.

Panevová J., 1980, *Formy a funkce ve stavbě české věty*. Academia, Praha.

Sgall P. et al.,1986, *Úvod do syntaxe a sémantiky. Některé nové směry v teoretické lingvistice*. Academia, Praha.

Šindlerová J., O. Bojar, 2009, Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Proceedings of TLT 8*, Milan, Italy, 185-196.