PDT-Vallex: trochu jiný valenční slovník

Zdeňka Urešová

uresova@ufal.mff.cuni.cz

Univerzita Karlova v Praze

Ústav formální a aplikované lingvistiky

Malostranské nám. 25

11800 Praha 1

Česká republika

Abstract

The Prague Dependency Treebank (PDT) contains as its integral part a valency lexicon called PDT-Vallex. PDT-Vallex lists (among other non-verbal entries) all the verbs that occur in the PDT and it also distinguishes their senses. A valency frame is formally described for each sense. This valency frame includes a list of verb complementations and their required surface form(s), for obligatory as well as optional arguments. Every occurrence of a verb in the PDT is linked to the appropriate verb sense entry in the PDT-Vallex lexicon. In this paper, we describe in detail the underlying approach to valency in the PDT, the format and contents of the PDT-Vallex entries and its relation to the Prague Dependency Treebank.

Keywords Computational Linguistics, Valency, Lexicon, Corpora, Linguistic Annotation, Dependency Syntax, Semantics, Tectogrammatical Representation, Functional Generative Description

Klíčová slova Počítačová lingvistika, valence, lexicon, korpusy, lingvistická anotace, závislostní syntax, sémantika, tektogramatická reprezentace, funkční generativní popis

1 Pozadí vzniku – teorie FGP a korpus PZK

Počítačový valenční slovník PDT-Vallex je budovaný na základě manuálně syntakticky anotovaného korpusu českých textů, nazvaného Pražský závislostní korpus (PZK). Zpracování valence v tomto slovníku vychází z principů valenční teorie FGP (Sgall et al., 1986; Panevová, 1974-5). Práce na slovníku umožnila konfrontaci teoretického zpracování valence s praxí. Z této skutečnosti vyplynula největší přednost slovníku, a sice jeho přístup ke zpracování valence. PDT-Vallex přistupuje k valenci „zdola“, tzn. slovník při tvorbě valenčních rámců a příkladů čerpal z reálných textů, z reálných korpusových dat.

Primárně měl slovník sloužit anotátorům k udržení konzistence při přiřazování funktorů slovesným doplněním. Po dokončení anotace PZK sloužil navíc ke kontrole anotace tektogramatické roviny PZK.

Slovník zachycuje jen ta slova (slovesa, substantiva, adjektiva a adverbia) a hlavně ty jejich významy, které se vyskytly v anotovaných datech. Celkem slovník obsahuje 10039 různých slov. Z tohoto počtu připadá 5510 hesel na slovesa, 3727 hesel na substantiva, 791 hesel na adjektiva a 11 hesel na adverbia. Celkový počet valenčních rámců ve slovníku je 14979.

2 Slovesná část slovníku

Povšimneme si základní, tj. slovesné části valenčního slovníku. Slovník obsahuje 5 510 sloves, která byla použita v 8 500 významech. V průběhu syntaktické anotace korpusu zpracovávali anotátoři primárně pouze ta slovesa, která se vyskytla v datech. Každý výskyt slovesa v PZK má tudíž odkaz na jeden z valenčních rámců ve slovníku PDT-Vallex (srov. Hajič et al., 2003, Hajič, Urešová, 2003).[1]

[1] Pro odkaz do valenčního slovníku slouží atribut val_frame.rf. Jeho hodnotou je identifikátor valenčního rámce, který je označeným uzlem (a jeho podstromem) realizován. T-lema v datech a t-lema ve slovníku se vždy shodují.

2.1 Valenční heslo a jeho obsah

Valenční heslo ve slovníku obsahuje:

1. t-lema - pro jedno t-lema může být ve slovníku více valenčních rámců. Každý valenční rámec v zásadě odpovídá jednomu významu, který může být buď konkrétní, abstraktní nebo frazeologický. Rozlišení těchto typů významů u různých sloves není ve slovníku konzistentní, je spíše intuitivní a není ani explicitně zachyceno.

2. valenční rámec - ve valenčním rámci je specifikace:

(a) Počtu členů rámce - počet členů je fixní (může být i nulový).
(b) Pojmenování členů rámce - členy rámce se rozlišují pouze pomocí funktorů
(c) Charakteristiky členů rámce - charakteristikou členu rozumíme v souladu s FGP rozlišení obligatornosti a fakultativnosti.
(d) Povrchové realizace (formy) - uvádí se pouze „základní" forma realizace rámce (tzv. „kanonický rámec"), která odpovídá aktivnímu užití slovesa. Kanonický rámec odpovídá základní diatezi a pro ostatní, tj. odvozené diateze se korespondence formy a rámce řeší prostřednictvím transformací.
(e) Příklady - příklad zachycuje nějaké konkrétní lexikální naplnění daného rámce. Jedná se o minimální srozumitelný fragment české věty, který obvykle pochází z textů PZK, ale někdy se jedná o nově vytvořený anebo upravený příklad. V případě, že může vzniknout pochybnost, které slovo z příkladu se vztahuje ke kterému členu rámce, je u takového slova v příkladu uveden i funktor.
(f) Poznámky - poznámky pomáhají významově rozlišovat mezi jednotlivými rámci uvnitř slovníkového hesla. Jako poznámka se používají pouze synonyma, synonymní víceslovné výrazy, antonyma nebo vidové protějšky. Nejsou na rozdíl od předchozího povinnou složkou hesla, avšak uvádějí se téměř vždy. Zřídka se uvádějí zejména u frazeologických významů, kde je význam zřejmý již z realizace rámce.

Jak již bylo řečeno, z valenčního rámce ve slovníku PDT-Vallex se uživatel dozví, kolik členů rámec obsahuje (počet), jak se jednotlivé členy jmenují (pojmenování) a

jakého jsou charakteru (typová charakteristika obligatornosti). Na rozdíl od některých tradičních valenčních přístupů (Daneš 1985, Pauliny 1943) nejsou ve valenčním rámci explicitně vyznačeny pozice valenčního doplnění, tj. valenční členy nejsou rozlišeny na levovalenční a pravovalenční.

2.1.1 Valenční rámec a význam

Valenční rámec zachycuje valenční doplnění daného slovesa. Jedno sloveso může nést více významů, a může tedy mít více valenčních rámců. Jeden valenční rámec odpovídá jednomu významu slovesa. S více významy počítáme u sloves, kde nás k tomu přímo vybízejí odlišná valenční doplnění:

Př. sloveso *přišít*:

přišít[1] (připevnit šitím) – *přišít knoflík na košili* – má tři členy rámce:

- ACT (kdo přišívá)
- PAT (co se přišívá)
- DIR3 (směr - kam se to přišívá)

přišít[2] (přeneseně: označit, přisoudit, uštědřit) – *přišít hráči pokutu* – má tři (jiné) členy rámce:

- ACT (kdo přišívá)
- PAT (co se přišívá – vlastnost, facka)
- ADDR (komu se to přišívá)

Jindy jsou naopak valenční doplnění ohodnocena stejně, ale významový rozdíl je zcela zjevný; i v tomto případě bude ve slovníku více rámců:

Př. sloveso *dělat*:

dělat[1] (být někým) – *dělat funkcionáře* – má dva členy rámce:

- ACT (kdo dělá)
- PAT (koho dělá)

dělat[2] (zabývat se) – *dělat politiku* – má dva (stejné) členy rámce:

- ACT (kdo dělá)
- PAT (co dělá)

Výjimečně může být více (např. Wordnetovských) významů (Fellbaum, 1998 i pro jeden rámec:

Př. *hučet v komíně* i *hučet v uchu* je jeden rámec pouze ACT a LOC.

Rozlišování polysémie (Čermák, 1995) tedy není (měřeno přísnými lexikografickými měřítky) v PDT-Vallexu zcela důsledné. Určování povahy a počtu významů pro daný lexém je velmi obtížné a hranice mezi jednotlivými významy je často těžko definovatelná.

Ve valenčním rámci se postihuje nepřímo rozdíl mezi konkrétním, abstraktním a frazeologickým užitím slovesa, i když explicitně typ významového rozdílu ve slovníku zachycen není. Nemáme významy rozlišeny indexy, jako se to značí v tradiční literatuře. PDT-Vallex rozlišuje konkrétní, abstraktní a frazeologický význam takto (srov. Mikulová et al., 2005, s. 106) :

- Konkrétní významy slovesa jsou takové významy daného slovesa, které přímo vyplývají z jeho lexikální sémantiky, jsou to jeho významy základní (původní), nepřenesené.
- Abstraktní významy slovesa jsou takové významy daného slovesa, které vznikají metaforickým (přeneseným) užitím významů konkrétních.
- Frazeologické významy nese dané sloveso tehdy, vystupuje-li jako součást nové víceslovné lexikální jednotky.

Ve slovníku je vedle významu základního a přeneseného naznačen ještě tzv. vyprázdněný význam, a to pro případy, kdy se význam celého spojení přesouvá na substantivum. Například ve spojení *podat stížnost* nese hlavní význam substantivum *stížnost* a sloveso *podat* má vyprázdněný význam.

Valenční rámec základního významu má ohodnoceny členy rámce základními běžnými funktory, kdežto pro rámce s přeneseným významem je kromě běžných funktorů u jednoslovných jednotek charakteristické užití funktoru

DPHR, a to pro víceslovné frazeologické jednotky. Pro rámce s vyprázdněným významem se používá výhradně funktor CPHR, naznačující specifický druh frazeologického významu.

Příklad slovesa *nést* se všemi rozlišovanými významy:

	Nést	
Význam	Rámec	Příklad
základní	ACT(.1) PAT(.4) ADDR (.3)	nést tatínkovi knihy
přenesený jednoslovný	ACT(.1) PAT(.4)	nést jméno
přenesený víceslovný (frazeologický)	ACT(.1) DPHR (kůže:S4, na-1[trh:S4]	nést kůži na trh
vyprázdněný	ACT(.1) CPHR {odpovědnost,…}.4	nést odpovědnost

V průběhu anotování PZK vznikala se vzrůstajícím množstvím anotovaných dat potřeba zavést další sémanticky vyhraněné funktory. Vzhledem k charakteru anotování však k jejich zavedení v praxi nedošlo. V teorii otevřené otázky bylo třeba vyřešit i za omezujícího předpokladu neměnit repertoár funktorů. Přesto některé podněty k zavedení nových funktorů nepřišly vniveč a byly alespoň teoreticky vzaty v úvahu (např. Lopatková a Panevová, 2005).

2.1.2 Počet členů valenčního rámce

Počet členů rámce je fixní. Většinou jeden člen rámce znamená také jeden příslušný funktor. Pouze v případech, kdy rámec obsahuje místo jednoho funktoru seznam alternujících funktorů, počet funktorů na jeden člen rámce může být vyšší. Počet členů rámce ale zůstává stálý.

V současné verzi slovníku existuje vzhledem k jemnému rozlišování například místních a časových významů více rámců právě kvůli fixnímu počtu členů rámce. Např. pro sloveso *umístit* je místní valenční doplnění buď DIR3 (směr) nebo LOC (místo): *umístit knihy do ústavu*.DIR3 nebo *umístit knihy v ústavu*. LOC. Sloveso *umístit* má dva

rámce: jeden se členem ohodnoceným DIR3 a druhý se členem ohodnoceným LOC. Valenční rámec totiž nesmí v současné verzi slovníku obsahovat alternativu tohoto druhu, kdy se na pozici jednoho členu valenčního určení místa „tlačí" dva funktory.

Počet členů rámce může být i nulový. To znamená, že rámec neobsahuje žádný valenční člen (žádný aktant ani žádné obligatorní volné doplnění). Mluvíme pak o „prázdném" valenčním rámci s notací EMPTY, např. pro sloveso *foukat, hřmět, lít*.

2.1.3 Pojmenování členů rámce

Jednotlivé členy valenčního rámce jsou označeny prostřednictvím funktorů. V jednom valenčním rámci nemohou být v souladu s teorií valence FGP dva stejně pojmenované členy (tj. stejné funktory). Členy mohou být označeny buď funktory pro aktanty, anebo funktory pro volná doplnění (srov. Mikulová et al., 2005). Člen valenčního rámce, který má charakter aktantu, ale zároveň není sémanticky samostatnou jednotkou, nýbrž závislou částí spojení, je pojmenován funktorem CPHR (Compound PHRaseme) pro tzv. složené predikáty a funktorem DPHR (Dependent PHRaseme) pro frazeologická spojení.

Tabulka funktorů pro aktanty:

Funktor	Plný název	Příklad
ACT	ACTor	Maminka.ACT vaří
PAT	PATient	Malovat obraz.PAT
ADDR	ADDResse	Darovala Mirce.ADDR knihu
EFF	EFFect	Přeložila publikaci do angličtiny.EFF
ORIG	ORIGin	Půjčil si peníze od kamaráda.ORIG

Tabulka funktorů pro valenční doplnění ve složených predikátech a ve frazeologických spojeních:

CPHR	Compound PHRaseme	Pokládal jim otázky.CPHR
DPHR	Dependent PHRaseme	Postavil si hlavu.DPHR

Funktorů pro volná slovesná doplnění je celkem 36 (srov. Mikulová et al., 2005, s. 427).

Funktory pro volná doplnění se objevují ve valenčním rámci jen v případě, že jsou obligatorní. Ne všechna vyjmenovaná volná doplnění se objevila jako obligatorní valenční doplnění. Nejčastější volné obligatorní doplnění v rámci je doplnění místa: LOC a DIR3.

Na pořadí členů ve valenčním rámci ve slovníku nezáleží, je dáno konvencí.

V některých případech je pojmenování členu rámce složeno z několika alternativních funktorů. Příkladem valenčního rámce s alternativními doplněními je valenční rámec pro jeden z významů slovesa *chovat se*.

chovat se: ACT(.1) MANN(*)|CRIT(*)|ACMP(*)|BEN(*)|CPR(*)

- *chová se laskavě*.MANN
- *ch. se podle pravidel*.CRIT
- *ch. se otrocky*.CPR
- *ch. se bezchybně*.ACMP
- *ch. se ku prospěchu věci*.BEN

Pokud je situace složitější a nebylo možné ji zachytit jednoduchou alternativou funktorů, je pro daný význam použito více rámců. Ve slovníku ale není nikde naznačeno, že tyto rámce spolu významově souvisejí.

2.1.4 Charakteristiky členů rámce

Valenční členy charakterizujeme v rámci z hlediska jejich obligatornosti či fakultativnosti. (O tomto kritériu srov. Panevová, 1974-75). V PDT-Vallexu se zapisují

do valenčního rámce pouze aktanty (jak obligatorní, tak fakultativní) a obligatorní volná doplnění.

2.1.5 Povrchové realizace (formy)

Valenční doplnění je ve valenčním rámci zapsáno jednak prostřednictvím funktoru, který zachycuje příslušný typ závislosti a jednak prostřednictvím povrchově-syntaktické realizace, která zachycuje slovnědruhové a morfematické vlastností valenčního doplnění.

Hodnoty funktorů jsou popsány výše, kap. 2.1.3. Otazník před zapsaným funktorem valenčního doplnění označuje fakultativnost, pokud před funktorem otazník není, označuje tento funktor obligatorní valenční doplnění.

Zápis valenčního rámce platí pro ty povrchové realizace (formy), které valenční doplnění získávají, je-li jejich řídící sloveso užito v aktivu. Povrchové realizace (formy), které valenční doplnění získávají, je-li jejich řídící sloveso užito v sekundárních diatezích, se do valenčního rámce nezapisují. Aby ale valenční rámec vyhovoval i těmto pravidelným formám, je ošetřen transformačními pravidly tak, že po jejich aplikaci lze daný valenční rámec použít i pro formy, které valenční doplnění získávají v sekundárních diatezích.

V PZK se na rozdíl od tradičního způsobu zápisu valenčního rámce používá rozšířený, formalizovaný zápis povrchových realizací jednotlivých členů rámce.

Typ závislosti se zapisuje hranatými závorkami a sesterské uzly se oddělují čárkou. Zápis znázorňující závislost vypadá následovně: řídící-uzel[závislý-uzel1,závislý-uzel2, závislý-uzel3].

Slovnědruhové a morfematické vlastnosti jednotlivých valenčních doplnění se zapisují ve velmi zkrácené formě za oddělovací symbol, kterým je tečka nebo dvojtečka. Přitom se dodržuje toto pořadí: slovní druh, rod, číslo, pád, stupeň a shoda u přídavných jmen. Pokud není některá z těchto kategorií v zápisu povrchově-syntaktické realizace uvedena, znamená to, že tato kategorie může nabývat pro dané valenční doplnění jakýchkoli hodnot. Slovní druh je zapsán malým písmenem, např. *a* přídavné jméno, *d* příslovce, *i* pro částice, *v* pro sloveso. Kořen přímé řeči je označen malým s, kořen takové

obsahové závislé klauze, která je uvozena vztažným zájmenem či příslovcem, je označen malým c. Rod je zapsán velkým písmenem: *F* ženský, *M* mužský životný, *I* mužský neživotný a *N* střední. Číslo je zapsáno velkými písmeny: *S* jednotné, *P* množné. Pád je zapsán svým číslem - 1 až 7. Některé z výše uvedených speciálnějších forem, např. stupeň a shoda, se většinou týkají idiomů (DPHR). Stupeň přídavného jména je označen (kvůli odlišení od pádu) symbolem @. Shoda v pádě, čísle a rodě s řídícím uzlem je označena symbolem # (pouze v případě, že tato kategorie u obou uzlů existuje a u závislého uzlu už to není zápisem morfologických požadavku specifikováno).

Pokud výše uvedené zápisy nestačí k popisu morfologických požadavků, je možné další požadavky uvést formou výčtu hodnot povolených na konkrétních pozicích morfologické značky. Zápis požadavku na hodnotu určité pozice morfologické značky začíná symbolem $, za ním je uvedeno číslo pozice (1 až 15) a následuje řetězec uzavřený do špičatých závorek < >. Tento řetězec je tvořený právě všemi znaky, které zápis na dané pozici morfologické značky povoluje. Všechny znaky (kromě písmen, číslic a spojovníku) vyskytující se uvnitř špičatých závorek musejí být uvozeny zpětným lomítkem. Vlnka (~) značí požadavek na to, aby v morfologické značce byl přítomen příznak pro negaci. Pokud je valenční rámec prázdný, je zapsán jako: *EMPTY*.

Příklady zápisu povrchově-syntaktických realizací:

Př.1 Specifikace (pouze) pádu: .4

Př. 2 Předložka a pád: s[.7]

Př. 3 Předložka a pád nebo samotný pád: pro[.4];.3

Př. 4 Závislá klauze (kořen je sloveso) uvozená podřadicí spojkou *že* nebo *aby*: že[.v];aby[.v]

U obligatorních volných doplnění je zápis povrchově-syntaktické realizace vynechán, pokud se dané valenční doplnění vyjadřuje prostředky obvyklými pro daný

funktor. Namísto explicitního zápisu povrchově syntaktické realizace je pak v závorce za funktorem hvězdička. U aktantů je však povrchově-syntaktická realizace uvedena vždy.

Příklady zápisu rámců:

Př. 1 Tranzitivní sloveso: ACT(.1) PAT(.4)

Př. 2 Infinitiv: ACT(.1) PAT(.f)

Př. 3 Frazém: ACT(.3) DPHR(mráz.S1,po-1[záda:P6])

Př. 4 Rámec s fakultativním členem: ACT(.1) PAT(.4) ?ORIG(z-1[.2]) ?EFF(na-1[.4])

3 PDT-Vallex a návaznost na data PZK

Slovník PDT-Vallex, budovaný souběžně s anotací Pražského závislostního korpusu, je zpracován tak, že ve valenčním rámci se uvádějí pouze ty povrchově-syntaktické realizace, které daná valenční doplnění mají při užití slovesa v aktivním tvaru (jak je to pro valenční slovníky běžné). Pokud jsou však v korpusu použity sekundární diateze, forma vyjádření valenčních doplnění neodpovídá kanonické formě jejich vyjádření uvedené v přiřazeném valenčním rámci. Valenční rámec v zápisu slovnědruhových a morfematických vlastností pravidelné změny forem neobsahuje, ale díky tzv. transformačním pravidlům jsou valenční rámce PDT-Vallexu aplikovatelné i na tyto transformované formy sekundárních diatezí. Zejména formy pro sekundární diateze, jako je pasívum reflexivní, pasívum opisné, rezultativ (mít + participium, dostat + participium), dispoziční modalita a reciprocita, jsou ošetřeny transformačními pravidly, jejichž aplikace tudíž umožnila přiřazení valenčního rámce ke každému výskytu všech sloves uvedeného korpusu.

4 Závěr

Jak již bylo řečeno, PDT-Vallex sice vznikal nejprve jako vedlejší produkt syntaktické anotace, ale postupně se stal důležitým zdrojem jak dalšího lingvistického výzkumu, tak počítačového zpracování češtiny. Tento valenční slovník je veřejně přístupný jako součást PZK (verze 2), který byl vydán v Linguistic Data Consortium (http://www.ldc.upenn.edu, LDC2006T01).

Výzkum popsaný v tomto článku byl podporován projekty GAUK 52408/2008, ME09008 a MSM0021620838.

Literatura:

Čermák, F. (1995): Manuál lexikografie. 283 stran. Praha.

Daneš, F. (1985): Věta a text. Studie ze syntaxe spisovné češtiny. Academia, Praha.

Fellbaum, Ch. (1998): WordNet: An Electronic Lexical Database. 445 stran. Cambridge, MA and London. MIT Press.

Hajič J., Urešová Z. (2003): Linguistic Annotation: from Links to Cross-Layer Lexicons**.** In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pp. 69-80. Vaxjo University Press.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V. (2003): PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation**.** In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pp. 57-68. Vaxjo University Press.

Lopatková,M., Panevová, J. (2005): Recent developments of the theory of valency in the light of the Prague Dependency Treebank. In: Insight into Slovak and Czech Corpus Linguistic. Mária Šimková. Str. 83-92. Veda Bratislava, Slovakia.

Mikulová et al. (2005): Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. TR-2005-28. 1185 stran. Prague. ÚFAL MFF UK, Prague.

Panevová, J. (1974-75): On Verbal Frames in Functional generative Description. Part I, The Prague Bulletin of Mathematical Linguistics 22, pp.3-40, Part II, The Prague Bulletin of Mathematical Linguistics 23, pp. 17-52

Pauliny, E. (1943): Štruktúra slovenského slovesa, SAVU, Bratislava

Sgall, P., Hajičová, E., Panevová, J. (1986): The Meaning of the Sentence in Its Semantic and Pragmatic Aspects (ed. by J. Mey), Dordrecht: Reidel and Prague: Academia.