Testování konzistence a úplnosti valenčního slovníku českých sloves*

Markéta Lopatková and Zdeněk Žabokrtský

Center for Computational Linguistics, MFF UK, Prague
{lopatkova,zabokrtsky}@ckl.mff.cuni.cz

Introduction

Valence je jeden ze základních jazykových jevů, se kterým je třeba počítat při tvorbě většiny aplikací v oblasti počítačového zpracování přirozeného jazyka a jehož zkoumání je zajímavé i pro "tradičního" lingvistu. Valenční vlastnosti sloves (i některých ostatních slovních druhů) jsou ovšem velmi rozmanité. Nelze je odvodit obecnými pravidly, je třeba je popsat v podobě valenčního slovníku, který obsahuje popis valence jednoho slova po druhém. Z těchto důvodů vzniká v Centru komputační lingvistiky od roku 2001 elektronický valenční slovník českých sloves VALLEX.[1] V tuto chvílu je v něm obsaženo zhruba 1400 sloves, probíhá jeho další rozšiřování.

Na moderní valenční slovník je kladena řadu požadavků - kromě strojové čitelnosti a dostatečné explicitnosti použitého popisu jde zejména o kvalitu dat ve slovníku obsažených. Slovník by neměl obsahovat chyby, a to ani z technického, ani z lingvistického úhlu pohledu. Mezi měřítka kvality slovníku řadíme konzistenci (důsledné zachycování "stejných věcí stejně") a úplnost (pokrytí všech významů, kterých dané sloveso může v jazyce nabývat).

V sekci 1 přiblížíme základy použité podkladové teorie, valenční teorii Funkčního generativního popisu češtiny. Dále popíšeme strukturu hesel slovníku VALLEX (sekce 2). Jádro článku tvoří sekce 3 a 4, ve kterých přibližujeme navržené nástroje (sekce 3) a rozebíráme metody testování kvality slovníku (sekce 4) - zejména porovnávání s existujícími zdroji (4.2), testování "vnitřní" konzistence slovníku (4.3) a ověřování na autentických větách (4.4). V sekci 5 uvedeme příklady aplikací, ve kterých se komplexní valenční slovník VALLEX s úspěchem využívá.

* Valenční slovník českých sloves VALLEX je vytvářen v Centru komputační lingvistiky při MFF UK, které vzniklo jako výzkumné centrum LN00A063 na základě programu MŠMT ČR.

[1] Vytváření slovníku VALLEX je spojeno s budováním Pražského závislostního korpusu (PDT), konkrétně s anotací na tektogramatické rovině, pro jejíž potřeby je vytvářena větev PDT-VALLEX (slouží ke konzistenci anotace valence v PDT, viz [1]).

1 Trocha teorie - co je valence?

Pokud aspirujeme na vytvoření konzistentního jazykového zdroje (language resource), který by byl využitelný pro aplikace v NLP i pro podrobná lingvistická zkoumání, potřebujeme důkladně rozpracovanou podkladovou teorii. VALLEX je budován na základě Funkčního generativního popisu (FGD) [6], což je závislostně orientovaný stratifikační systém, v jehož rámci je teorie valence studována od sedmdesátých let (viz zejména [5]).

Co je to tedy valence? Podle autorů valenčního slovníku Slovesa pro praxi [7]:[2] *"Valencí rozumíme v lingvistice schopnost lexikální jednotky, především slovesa, vázat na sebe jiné výrazy a mj. tak zakládat větné struktury."* Tato schopnost se týká primárně významové reprezentace, promítá se i do povrchové realizace věty.

Informace o valenčním chování lexikální jednotky je uchovávána ve valenčních rámcích - každému slovesu odpovídá soubor **valenčních rámců**, které ve FGD v zásadě odpovídají jednotlivým významů slovesa. Valenční rámec se skládá z **vnitřních doplnění slovesa** (aktantů, též participantů nebo argumentů), obligatorních i fakultativních, a dále z **obligatorních volných doplnění** (adverbiální doplnění, adjunkty).

FGD rozlišuje pět vnitřních doplnění (Actor, Patient, Addressee, Origin, Effect; v aktivní větě Actor typicky odpovídá subjektu, Patient přímému objektu, Addressee nepřímému objektu) a řadu volných doplnění (odpovídají příslovečným určením, např. místa, času, způsobu, prostředku, podmínky; seznam viz [2]). Vnitřní i volná doplnění mohou být buď obligatorní (povinně přítomny ve významové reprezentaci věty),[3] nebo fakultativní.[4]

*Matka.*ACT *předělala dětem.*ADDR *loutku.*PAT *z Kašpárka.*ORIG *na čerta.*EFF.
*Petr.*ACT *včera.TWHEN v novinách.LOC četl o katastrofě.*PAT.
*Děti.*ACT *přišli pozdě.TWHEN.* (=*domů, sem.*DIR1)
*Venku.*LOC *prší.*
V Praze.LOC se sejdeme na Hlavním nádraží.LOC u pokladen.LOC.
*Kniha.*ACT *vyšla.*
*Chlapec.*ACT *vyrostl v muže.*PAT.

Klasifikaci FGD obohacujeme o tzv. typická doplnění.[5]), z nichž některá mohou být obligatorní (*přijít kam.*DIR3).

[2] Tento slovník je cenným zdrojem dat, obsahuje ovšem jen omezený počet sloves (767 sloves) a není strukturován tak, aby bylo možné využít ho v automatických procedurách (více viz sekce (4.2)).

[3] Jako takové ovšem mohou být vypuštěny v povrchové realizaci věty - mohou být známy z např. kontextu.

[4] Následující příklady částečně přebíráme z článků prof. Panevové. Členy valenčních rámců sloves jsou označeny verzálkami; fakultativní volná doplnění, která nejsou součástí rámce, označujeme kurzívou (přesněji - vyznačujeme jméno příslušné sémantické relace mezi slovesem a jeho valenčním doplněním).

[5] Typická doplnění jsou fakultativní volné doplnění (tudíž nepatřící do "klasického" valenčního rámce), které dané sloveso "zpravidla" rozvíjí; navíc takové doplnění obvykle rozvíjí celou třídu sémanticky blízkých sloves. Např. slovesa pohybu jsou typ-

2 Co valenční slovník obsahuje?

Každé sloveso ve slovníku VALLEX je reprezentováno jako soubor valenčních rámců s doplňujícími syntakticko-sémantickými informacemi (vztaženými vždy k danému rámci); homonymní slovesa jsou popsána více soubory. Typicky jeden rámec odpovídá jednomu významu slovesa, příslušný význam je vždy určen glosou a příklady použití. Valenční rámec slovesa, který tvoří jádro zachycované informace, definujeme jako kombinaci prvků rámce (slovesných doplnění). U každého prvku rámce jsou zachyceny jeho tři vlastnosti:

– funktor, tj. jméno sémantické relace mezi slovesem a jeho příslušným doplněním (aktantem nebo volným doplněním);
– morfematické vyjádření příslušného doplnění (číslo pádu, předložka+číslo pádu, infinitiv nebo podřadící spojka);
– typ doplnění, tj. zda jde o obligatorní (obl) nebo fakultativní (opt) valenční doplnění, příp. doplnění typické (typ).

Cílem VALLEXu je poskytnout uživateli komplexní syntakticko-sémantickou informaci. Proto je jádro slovníku - soubor valenčních rámců - obohaceno o další informace využitelné v NLP (tyto údaje jsou vždy vztaženy k jednotlivým valenčním rámcům, nikoli k celému slovesu - výjimku tvoří vidová charakteristika, která je vlastní celému slovesu):[6]

– reflexivita* (výčet možných syntaktických funkcí zvratného zájmena *se/si*);
– reciprocita* (možnost členu valenčního rámce vstupovat do symetrické relace s jiným členem);
– kontrola (u sloves s doplněním ve formě infinitivu; jde o vzájemný vztah mezi některým členu valenčního rámce a subjektem infinitivu);
– vid a příp. vidový protějšek (odkaz na příslušný valenční rámec);
– syntakticko-sémantická třída*;
– pointer na odpovídající synset české větve sémantické databáze EuroWordNet.*

Ve valenčním slovníku VALLEX 1.0 je obsaženo přes 1400 českých sloves - prvních zhruba 1000 sloves bylo vybráno podle frekvence v Českém národním korpuse (s výjimkou pomocného slovesa *být*, které vyžaduje zvláštní zpra-cování), k nim byly posléze doplněny jejich vidové protějšky (pokud ještě nebyly zpracovány).

3 Jaké nástroje lze využít při testování konzistence a úplnosti VALLEXu?

Při budování slovníku je nutno klást maximální důraz na systematičnost a konzistenci v zachycování jednotlivých jazykových jevů, neboť konzistence zpracování patří k základním požadavkům kladeným na každý zdroj jazykových dat.

icky rozvíjena volnými doplněními směru (*jít jet/běžet/spěchat do kina*.DIR3/*přes les*.DIR2/*z domova*.DIR1

[6] Údaje označené hvězdičkou jsou zpracovány zatím jen částečně.

Přestože při testování konzistence slovníku mají a budou mít nezastupitelnou úlohu vzájemné ruční kontroly anotátorů (každé heslo procházejí nejméně tři lidé v různých fázích zpracování), jejich úsilí mohou podstatným způsobem zefektivnit navržené nástroje umožňující vyhledávání údajů a třídění hesel podle jednotlivých atributů a jejich kombinací.

3.1 Vyhledávací rozhraní pro WWW

Vyhledávací rozhraní pro WWW umožňuje vyhledávat rámce podle toho, zda daný rámec nebo jeho vybrané atributy obsahují určité podřetězce nebo odpovídají regulárnímu výrazu. (Např. "najdi všechna slovesa kontroly", "najdi všechna slovesa obsahující v rámci funktor EFF", "najdi všechna slovesa s reflexivním zájmenem *se*", případně "zobraz celý slovník" (dotaz bez omezovacích podmínek).)

Dále je možné zjišťovat rozvržení hodnot jednotlivých atributů. (Např. "zobraz všechny hodnoty atributu reciprocity a jejich rozložení", "zobraz valenční rámce všech sloves kontroly"). K vyhledaným hodnotám lze vždy zobrazit informaci o příslušných valenčních rámcích, případně o jejich vybraných atributech.

Toto rozhraní je grafické, umožňuje klást dotazy anotátorům, kteří nejsou zběhlí v programování.

3.2 Vyhledávání v dostupných elektronických zdrojích

Tato aplikace umožňuje rychle nahlédnout, jak je dané sloveso zpracováno v existujících slovnících. K dispozici máme slovníky Slovesa pro praxi a Slovník spisovného jazyka českého, dále případné zpracování slovesa v české větvi EuroWordNetu a 100 náhodných výskytů v Českém národním korpusu.

3.3 XSH

Datová reprezentace slovníku je založená na XML, lze tedy využít řady existujících nástrojů. Jde zejména o editor XSH (*XML Editing Shell*)[7] P. Pajase, který umožňuje klást dotazy přesahující možnosti grafického rozhraní (např. "zjisti počet sloves / rámců / prvků v rámcích", "zobraz slovesa, která mají více než 5 rámců", "najdi primární reflexiva tantum"). Užívání XSH vyžaduje základní znalost XML technologií, více viz [3].

4 Jaké metody lze využít při testování konzistence a úplnosti VALLEXu

Testování konzistence a úplnosti slovníku je metodologicky i časově náročná činnost.[8] Neznáme obecně přijatou metodologii testování systematičnosti a konzis-

[7] http://xsh.sourceforge.net

[8] Hrubý odhad času vynaloženého na testování konzistence a úplnosti slovníku se pohybuje okolo 1/3 času věnovaného vytváření slovníku.

tence slovníku, která by byla dostatečně efektivní a komplexní a kterou bychom mohli přejmout, proto jsme byli nuceni vypracovat vlastní metody testování.

Testování konzistence bylo částečně provedeno po základním zpracování tisíce českých sloves, druhé kolo masivního testování (a následné opravy) proběhlo po zpracování všech 1400 sloves obsažených ve verzi slovníku VALLEX 1.0.

4.1 Odstranění technických nedostatků

Slovník VALLEX má striktně definovanou notaci, prohřešky proti ní (např. chybějící závorka) lze většinou nalézt automaticky. Dalším typem čistě technické chyby je překlep ve funktoru nebo použití neexistující morfématické formy (např. u+4 - předložka *u* se nepojí s akuzativem).

4.2 Porovnání s jinými lexikografickými zdroji

Již při základním zpracování sloves jsme využívali valenční slovník BRIEF a Slovník spisovného jazyka českého (SSJČ). Při následném testování jsme obsah slovníku VALLEX porovnávali s tím, jak jsou slovesa zpracována ve slovníku Slovesa pro praxi (SPP) a částečně i v české větvi databáze EuroWordNet (EWN). Toto porovnání bylo přínosné zejména pro vyčlenění jednotlivých významů zpracovávaných sloves a pro doplnění případných chybějících významů slovesa, přitom ovšem bylo potřeba brát v úvahu rozdílné přístupy uplatněné v jednotlivých zdrojích.

BRIEF. Valenční slovník povrchových realizací ve formátu BRIEF [4], který vznikl kompilací několika tištěných slovníků, především SSJČ, byl využit již při základním zpracování sloves zejména jako zdroj morfematických forem, které se pojí s jednotlivými slovesy.

Slovník spisovného jazyka českého. SSJČ a jeho elektronická podoba[9] sloužila jako základní zdroj informací o významech sloves. Vyčlenění jednotlivých významů sloves v SSJČ však neodpovídá jednotlivým valenčním rámcům (tuto zásadu jsme převzali z podkladové teorie FGD), proto bylo přepracováno s důrazem na syntaktická a sémanticko-syntaktická kritéria.

Obecně jsou v SSJČ významy členěny jemněji (např. *bát se*), existují ovšem i příklady opačné relace (např. *pocházet*), viz tabulky 1 a 2.

Významy v SSJČ označené za zastaralé nebyly ve VALLEXu zpracovávány.

Slovesa pro praxi. Slovník SSP poskytuje podrobné údaje o valenčním chování vybraných sloves (767 sloves), které byly využity při testování VALLEXu. Vyčlenění jednotlivých významů ovšem opět zcela neodpovídá kritériím přijatým ve VALLEXu - sporná je především možnost přiřazování konkrétních užití slovesa jednotlivým rámcům (viz např. pět významů slovesa *bát se*, viz tabulku 3).

[9] Aplikace GSlov, adresa ???

SSJČ – *bát se*	VALLEX – *bát se*
1. mít strach ∼ *byla sama doma a bála se* 2. mít strach něco udělat ∼ *bojí se jít za tmy do lesa* 3. mít strach z někoho/něčeho ∼ *bát se otce, samoty*	1. ACT (PAT), mít strach ∼ *bát se tmy/učitele* */aby se v labyrintu vyznal* */že bude pršet;* *bojí se létat*
4. mít starost, že někdo/něco je ohrožen(o) ∼ *bát se o otce, o výsledky své práce;* *bojím se, abych neupadl*	2. ACT PAT, obávat se o někoho/něco ∼ *bála se o syna*

Tabulka 1. Vyčlenění významů slovesa *bát se* v SSJČ a ve VALLEXu.

SSJČ – *pocházet*	VALLEX – *pocházet*
1. vzít původ, vznik; vzniknout, vzejít, povstat, zrodit se	1. ACT PAT, ∼ *nemoc pochází z viru* 2. ACT DIR1, ∼ *Jan pochází z venkova* 3. ACT TFRWH, ∼ *rukopis pochází z roku 1352*

Tabulka 2. Vyčlenění významů slovesa *pocházet* v SSJČ a ve VALLEXu.

SPP – *bát se*
1. mít pocit ohrožení ∼ *Když ten pes pozná, že se ho bojíš, kousne tě docela určitě. Koně se báli biče jako čert kříže.* 2. mít obavu z něj. vlastní činnosti ∼ *Nakonec se našel nakladatel, který se nebál český překlad vydat. Z chlapce se stává muž. Nebojí se žádné práce.* 3. mít nelibý pocit plynoucí z očekávání něčeho nepříjemného ∼ *Hlavně se bojím toho, že budu nemocná. Psi zalezli do boudy, báli se, že je tentokrát výprask nemine. Ponejvíc se lidé bojí, aby je někdo neošidil.* 4. mít obavu o někoho, něco ∼ *O výsledky své práce se nebojíme. Bezpečnostní situace v hlavním městě je taková, že se obyvatelé právem bojí o svůj majetek a někteří i o své životy.* 5. být bojácný ∼ *Pojď, neboj se, nejsi přece malé dítě. Míša se nebojí, jaképak bání! Co je to za hlídacího psa, když se bojí!*

Tabulka 3. Vyčlenění významů slovesa *bát se* v SPP.

EuroWordNet. EuroWordNet[10] je multilinguální lexikální databáze; průnik sloves v její české větvi (cca 3 000 sloves) a sloves zpracovaných ve VALLEXu představuje zhruba 500 sloves. EWN neobsahuje žádné informace o valenci, pokusili jsme se jej částečně využít jako pomůcku pro rozlišování významů slovesa (s plným vědomím, že členění významů v EWN, jehož základem je zpracování anglických sloves, zcela neodpovídá češtině). Nicméně výhody i nedokonalého navázání jednotlivých valenčních rámců na synsety (tj. základní významové jednotky EWN) jsou zřejmé.

[10] http://www.hum.uva.nl/∼ewn/

4.3 Testování konzistence uvnitř VALLEXu

Mezi hlavní měřítka kvality slovníku je potřeba řadit konzistenci zpracování dat, nutnost stanovit jasnou koncepci (která může být pro různé účely různá) a v jejím rámci zpracovávat "stejné věci stejně". Proto je ve VALLEXu kladen velký důraz na odstranění nezdůvodnitelné různorodosti, která vzniká při budování slovníku "zdola".

Vidové protějšky. Valenční rámce sobě odpovídajících vidových protějšků jsou často totožné. Protože vidové protějšky byly zpracovávány nezávisle na sobě, lze jejich porovnání (a případné následné sjednocení) považovat za masivní test konzistence zpracování. Stejně jsme ve VALLEXu využili podobnosti předponových a bezpředponových sloves.

Sjednocení anotace příbuzných sloves je přínosné zejména pro slovesa s mnoha významy - např. vidové protějšky *brát* a *vzít* mají 13 totožných rámců zachycujících primární a posunuté užití a 9 totožných rámců pro idiomy, *brát* má navíc 4 idiomatické rámce, *vzít* rámce 2.

Sémantické třídy. Slovník VALLEX obsahuje u přibližně jedné třetiny rámců informaci o syntakticko-sémantické třídě. I když jde zatím pouze o předběžné třídění a seskupení slovesných rámců, má velký význam pro konzistenci zpracování - předpokladem je, že slovesa patřící do jedné třídy se budou chovat i z pohledu valence velmi podobně. Zatím byla systematicky provedena anotace u sloves pohybu (třídy motion, transport), sloves pravení (třídy communication, mental action, perception, social interaction) a částečně u sloves výměny (exchange).

Tímto způsobem bylo například u sloves pohybu (motion, transport) systematicky doplněno typické doplnění záměru (funktor INTT) vždy k primárnímu významu zpracovaných sloves, *jít na houby, přivedl mu ukázat přítelkyni* (původně 24 intuitivně anotovaných výskytů INTT bylo rozšířeno na 48 výskytů).

Morfématické formy. Systematicky byly zpracovány některé morfématické formy - byly porovnány všechny funktory s konkrétními formami i celé valenční rámce. Tyto testy byly účelné zejména pro zpracování předložkových skupin *o*+4 (zejména s ohledem na zachycení funktorů DIFF (difference, rozdíl) a OBST (obstacle, překážka)) a *za*+4 (systematické zpracování sloves výměny). Dále byla zkoumána doplnění vyjádřená infinitivem a výrazem *jako* (konzistentní rozlišování funktorů COMPL (complement, doplněk) a EFF (effect)).

Kromě toho byly zkoumány možné kombinace morfématických forem u jednotlivých funktorů (např. u funktoru INTT (intence, záměr) u sloves pohybu byla forma sjednocena na *na*+4, inf).[11]

Další možností je porovnávat kombinace forem bez ohledu na funktor (zejména např. pro úplný soubor podřadících spojek, zatím zpracováno částečně).

[11] Výjimku tvoří sloveso *nést* ve významu *nese rozdat handouty*, kde není možná předložková skupina *na*+4 (nejednotnost je tedy v tomto případě opodstatněná).

Typická doplnění. Systematicky jsou zpracovávána též fakultativní volná doplnění, která lze označit jako typická (viz poznámka 5). Byly porovnány všechny rámce, ve kterých se vyskytuje některé ze specifických volných doplnění (např. MEANS, BEN, CAUS).

Typická doplnění byla sjednocena také u sloves již zpracovaných sémantických tříd. Například slovesa pohybu (třídy motion a transport) jsou typicky rozvíjena volnými doplněními směru - pro určení obligatorního doplnění směru dává kritéria FGD, fakultativní doplnění jsou zpracována systematicky v rámci tříd; slovesa vyjadřující pohyb pomocí dopravního prostředku mají typicky MEANS (means, prostředek).

Četnost. Jako obecně užitečná se ukázala technika "co je málo časté, to je podezřelé". Tuto techniku lze s výhodou využít napříč slovníkem, u všech zachycovaných informací. Například u morfématické formy lze tímto způsobem odhalit nejen překlepy, ale i idiomatičnost některých spojení. U funktorů, které se ve VALLEXu vyskytly pouze několikrát, je potřeba zkontrolovat jejich účelnost, případně správné rozlišování anotátory (konkrétně např. funktory NORM, norma a CRIT, kritérium). Také ověřování anotace kontroly a reciprocity vedlo k omezení neopodstatněné různorodosti (málo četné hodnoty v těchto atributech vedly k odhalení technických nedostatků i faktických chyb).

Technika "co je málo časté, to je podezřelé" byla (zatím částečně) použita i na celé valenční rámce - pokud se některý rámec vyskytne v celém VALLEXu jen jednou, je vhodné ověřit, zda se v něm nevyskytuje nějaká chyba nebo neopodstatněná variace.

4.4 Ověřování na Českém národním korpusu

Zpracování sloves ve VALLEXu je ověřováno na autentických příkladech užití slovesa v ČNK.[12] Pro každé zpracované sloveso jsme použili 60 (pro nejsložitější slovesa 100) náhodně vybraných příkladových vět[13] z ČNK a ověřovali, zda lze výskytům daného slovesa přiřadit valenční rámec z VALLEXu. Přínosem této metody je především ověření vhodného rozčlenění slovesných rámců - důležitým kritériem pro vyčlenění jednotlivých valenčních rámců je shoda anotátorů v jejich přiřazování konkrétním výskytům slovesa -, případně doplnění chybějících rámců.

Například pro sloveso *nalézat*^{I} byly původně vyčleněny 4 rámce - 1. hledáním získávat, objevovat (*nalézat zlato na Aljašce*), 2. získávat (*nalézat přítele, potěšení v práci, pochopení*), 3. odhalovat (*nalézat na studiu kladné stránky*), 4. ohodnotit (*nenalézal na něm nic dobrého*); testy na příkladech ukázaly nemožnost rozlišovat mezi 2. a 3. rámcem, proto byly tyto dva rámce sloučeny (v souladu se SSJČ).

[12] http://ucnk.ff.cuni.cz

[13] Pro časovou náročnost těchto kontrol (60 x 1000 vět = 60 000 přiřazených výskytů valenčního rámce) bylo zatím použito pouze omezeného vzorku ČNK, předpokládáme další ověřování.

Naopak, na základě vět z ČNK byl pro sloveso *přijmout* vyčleněn nový rámec s glosou *schválit* (*parlament přijal zákon*).

5 K čemu valenční slovník?

Při budování VALLEXu je kladen důraz na skutečnost, aby byl slovník snadno a rychle čitelný, na snadnou orientaci a na srozumitelnost. To jsou základní předpoklady, které jsou nezbytné pro efektivní manuální zpracovávání jednotlivých sloves a pro možnost odhalování chyb a nekonzistencí. Na druhou stranu je takový formát podmínkou pro využití slovníku v dalším lingvistickém výzkumu. Nicméně hlavní přínos VALLEXu se předpokládá v automatických procedurách NLP.

V současné době se VALLEX testuje v následujících aplikacích:

– automatická syntaktická analýza ("shallow parsing");
– "tektogrammatický parser", tj. automatický systém pro vytváření významové reprezentace české věty;
– zdrojová data pro budování valenčního slovníku substantiv.

Valenční slovník VALLEX je pro nekomerční účely volně k dispozici, více informací viz http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/.

6 Shrnutí a otevřené otázky

Vytváření valenčního slovníku českých sloves VALLEX je úzce spojeno s budováním Pražského závislostního korpusu, jeho koncept vznikl v souvislosti s potřebou zajistit konzistentní zachycení valence v PDT. Zásadní důraz je přitom kladen na systematičnost zpracování všech jevů ve slovníku obsažených.

V tomto příspěvku jsme představili nástroje pro vyhledávání údajů a třídění hesel podle jednotlivých atributů, které byly navrženy pro testování konzistence a úplnosti slovníku. Dále jsme přiblížili řadu metod již použitých i v současné době aplikovaných - tyto metody jednak využívají existující jazykové zdroje, jednak se soustřeďují na eliminaci neopodstatněné různorodosti a na dosažení jednotného zpracování jevů ve slovníku obsažených.

Metody zde stručně popsané je možno chápat jako příspěvek k vytváření metodologie testování konzistence a úplnosti jazykových zdrojů. Zatím otevřenou otázkou zůstává metodologie evaluace slovníku, kvalifikovaný odhad možného množství chyb a mezianotátorské shody.

Reference

1. Hajič, J. 2003. ???. In: ???
2. Hajičová, E., Panevová, J., Sgall, P. 2001. Manuálů pro tektogramatické značkování. ÚFAL/CKL TR-2001-???

3. Lopatková, M., Žabokrtský, Z., Skwarska, K., Benešová, V. 2002. Tektogramaticky anotovaný valenční slovník českých sloves. ÚFAL/CKL TR-2002-15.
4. Pala K., Ševeček, P. 1997. Valence českých sloves (Valency of Czech verbs). In: *Sborník prací FFBU*. volume A45.
5. Panevová, J. 1994. Valency Frames and the Meaning of the Sentence. In: Ph. L. Luelsdorff (ed.) *The Prague School of Structural and Functional Linguistics*. Amsterdam-Philadelphia, John Benjamins, pp. 223-243.
6. 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.
7. Svozilová, N., Prouzová, H., Jirsová, A. 1997. Slovesa pro praxi. Academia, Praha.
8. Slovník spisovného jazyka českého. Praha. 1964.