Automatické přiřazování valenčních rámců a jejich slévání

Eduard Bejček

Univerzita Karlova v Praze,
Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta
Malostranské náměstí 25
11800 Praha
Česká republika
bejcek@ufal.mff.cuni.cz

Abstrakt *Informace o valenci sloves je podstatná pro mnoho odvětví NLP. Existuje proto již několik valenčních slovníků. V tomto článku představíme dva z nich (VALLEX a PDT-VALLEX), které jsou k disposici v elektronické podobě a které mají společné východisko. Oba mají své přednosti a naším cílem je spojit je v jeden slovník.*
Máme k disposici data z korpusu PDT, kterým jsou ručně přiřazeny položky prvního ze slovníků. To nám pomůže provázat oba slovníky přes data využívajíce automatické identifikace (následované ruční prací s problémovými případy). Tímto poloautomatickým spojením dvou slovníků vznikne kvalitní lexikografický zdroj, který by jinak vyžadoval mnohem více lidské práce.
Průměrná úspěšnost namapování rámce z jednoho slovníku výběrem náhodného rámce z druhého je přibližně 60 %.
Článek se také zmiňuje o universálním formátu, ve kterém bude výhodné nová data ukládat odděleně od stávajících.

Úvod

Sloveso je v lingvistice tradičně chápáno jako centrum věty a jako takové tvoří kořen stromu syntaktické analýzy. Všechny ostatní části věty jsou pak vhodně zavěšeny na slovese, nebo na některém následníku slovesa ve stromové hierarchii. Dobrý valenční slovník tedy usnadňuje parsing věty, neboť nabízí možná slovesná doplnění – tedy sloty, které závisí na slovese a mohou být naplněny dalšími částmi věty. Další využití valenčního slovníku spočívá v jeho schopnosti rozlišení významu: rozpoznáním valenčních doplnění můžeme s pomocí valenčního slovníku určit, o který z významů slovesa se jedná, a použít tak například při překladu vhodný ekvivalent v cizím jazyce, či zpřesnit výsledky úlohy „information retrieval“ atp. Pokud se nám podaří využít stávajících valenčních slovníků a s vynaložením menšího úsilí je spojit v jeden, získáme lexikografický zdroj cenný pro mnohá odvětví NLP (Natural Language Processing).

Cílem tohoto článku je představit probíhající práce na automatickém slévání dvou valenčních slovníků. V první kapitole popíšeme oba slovníky a zmíníme jejich výhody. V druhé kapitole čtenáře seznámíme s korpusem PDT, který využijeme pro slévání, a zastavíme se u formátu, který navrhujeme pro uchování dat. Předposlední kapitola přináší stručný rozbor velikosti slovníků, které máme k disposici, a poslední představuje práci na slévání slovníků ve dvou fázích.

1 Valence sloves a valenční slovníky

Valencí slovesa[1] se označuje jeho schopnost vázat další konkrétní syntaktické prvky věty. Některé z nich jsou vyžadovány (*obligatorní doplnění*), jiné jen povoleny (*fakultativní doplnění*). Sloveso může mít více než jeden tzv. *valenční rámec*; různé rámce jednoho slovesa obvykle representují různé významy tohoto slovesa. Oba slovníky, kterými se zabýváme, vycházejí z teorie valence Funkčního generativního popisu (FGP) češtiny ([4,5,6]). Ten dále klasifikuje typy doplnění. Rozlišuje tak například posice *actor, paciens, adresát, efekt, překážka, směr „odkud“, čas „do kdy“* apod.

Příkladem může být třeba sloveso *čekat* a jeho tři valenční rámce:

1. Petr$_{\mathsf{ACT}}$ **čekal** od Martina$_{\mathsf{ORIG}}$ omluvu$_{\mathsf{PAT}}$.
2. Petr$_{\mathsf{ACT}}$ **čekal** na Martina$_{\mathsf{PAT}}$ s večeří$_{\mathsf{ACMP}}$.
3. Petr$_{\mathsf{ACT}}$ **čekal** Martinovi$_{\mathsf{BEN}}$ s dluhem$_{\mathsf{PAT}}$.

První má význam „očekávat“, druhý „odložit (nějakou činnost)“ a třetí „počkat (obvykle právě s dluhem)“ – zde tedy platí, že s různými valenčními rámci se liší i významy slovesa. U prvního jsou všechna tři doplnění obligatorní, u druhého nikoli (lze říci „Petr čekal na Martina.“, ale nikoli *„Petr čekal od Martina.“). Také předložky a pády jednotlivých doplnění jsou jasně určeny. (Tj. začíná-li věta „Petr čekal na...“ a následuje akusativ (tedy ne třeba „Petr čekal na nádraží“), je tím jednoznačně určen valenční rámec 2 a tedy použitý význam slovesa *čekat* (tj. „odložit“)).

Valenční vlastnosti se sloveso od slovesa liší a nelze je odvozovat obecnými pravidly; proto je nutné je popsat pro každé sloveso zvlášť. Cílem našeho projektu je nový valenční slovník, který bude obsahovat jednak rámce, které se vyskytovaly v obou slovnících, sloučené

[1] Valenci mohou mít také některá slovesná substantiva a adjektiva. Těmi se v tomto textu až na výjimky zabývat nebudeme.

do jednoho a také ostatní rámce upravené, aby odpovídaly požadavkům na rámce výsledného slovníku.

Nyní se seznámíme se těmito dvěma existujícími elektronickými valenčními slovníky. V mnohém se podobají, ale každý má oproti druhému i své přednosti.

1.1 VALLEX

Slovník VALLEX ([3]) se začal vytvářet v roce 2001. Stejně jako slovník PDT-VALLEX vychází z FGP (viz [7]). Popisuje valenční vlastnosti 2730 lexémů[2] (vidové dvojice jsou obvykle popisovány dohromady; počítaje je zvlášť, dostali bychom přibližně 4500 sloves).

Pro každý takový lexém jsou uvedeny všechny jeho známé valenční rámce. Každý rámec obsahuje obligatorní a fakultativní doplnění, způsob jejich povrchového vyjádření (předložka a pád, či infinitiv, nebo podřadná spojka pro celou vedlejší větu), glosu a příklad usnadňující pochopení významu či použití tohoto rámce a často další doplňující syntaktické informace (reflexivita, reciprocita, kontrola a syntaktickosémantická třída). V záhlaví lexému jsou vyjmenovány jeho vidové varianty (navíc u jednotlivých valenčních rámců mohou být potom některé z nich explicitně vyloučeny). Ukázka lexému *odpovídat* je na obrázku 1.

Mezi výhody VALLEXu patří:

Komplexnost pro sloveso. Slovesa, která jsou ve slovníku zahrnuta, jsou zpracována plně, neměl by chybět žádný rámec; zahrnuta byla i častá idiomatická použití.

Výběr sloves. Slovník pokrývá slovesná lemmata, která se vyskytují v českém textu nejčastěji.

Bohatost informací. Slovník poskytuje glosu, informaci o vidu, reflexivitě, atd.

Pečlivé ruční zpracování. Slovník vznikal dlouho, každý lexém byl zkontrolován více lidmi a porovnáván s jinými existujícími slovníky. Dá se teda rozhodně označit za spolehlivý zdroj.

1.2 PDT-VALLEX

PDT-VALLEX ([1]) začal vznikat zároveň s tektogramatickou rovinou PDT (viz níže) od roku 2000. Bylo potřeba mít valenční slovník, na nějž by mohly odkazovat všechna slovesa (a slovesná substantiva a adjektiva) z korpusu. Protože VALLEX ještě nebyl k dispozici, začaly se vyvíjet oba dva valenční slovníky paralelně.

Struktura (jak je vidět na obrázku 2 zobrazujícím opět sloveso *odpovídat*, tentokrát v PDT-VALLEXu) je podobná jako u VALLEXu. Opět je jedno lemma (tentorkát bez vidového protějšku) rozděleno na několik valenčních rámců. U každého jsou uvedeny charakteristiky jednotlivých doplnění (actor, paciens, atd. a pád, předložka, nebo spojka) a několik příkladů užití. Dále je uveden identifikátor rámce (pro provázání s PDT) a počet výskytů v PDT.

Výhody PDT-VALLEXu jsou zejména:

Anotace v datech. Velká deviza je skryta právě v jeho raison d'être: v jeho provázání s daty. Díky tomu jsou k disposici data pro kontrolu slovníku, pro trénování přiřazování rámců – a pro naši úlohu data, která napomohou provázání s jiným slovníkem.

Nejen slovesa. Slovník obsahuje kromě sloves (např. „Jan odpovídá Evě na dotaz“) také některá slovesná substantiva („odpovídání Jana Evě na dotaz“) a adjektiva („odpovídající na dotaz“).

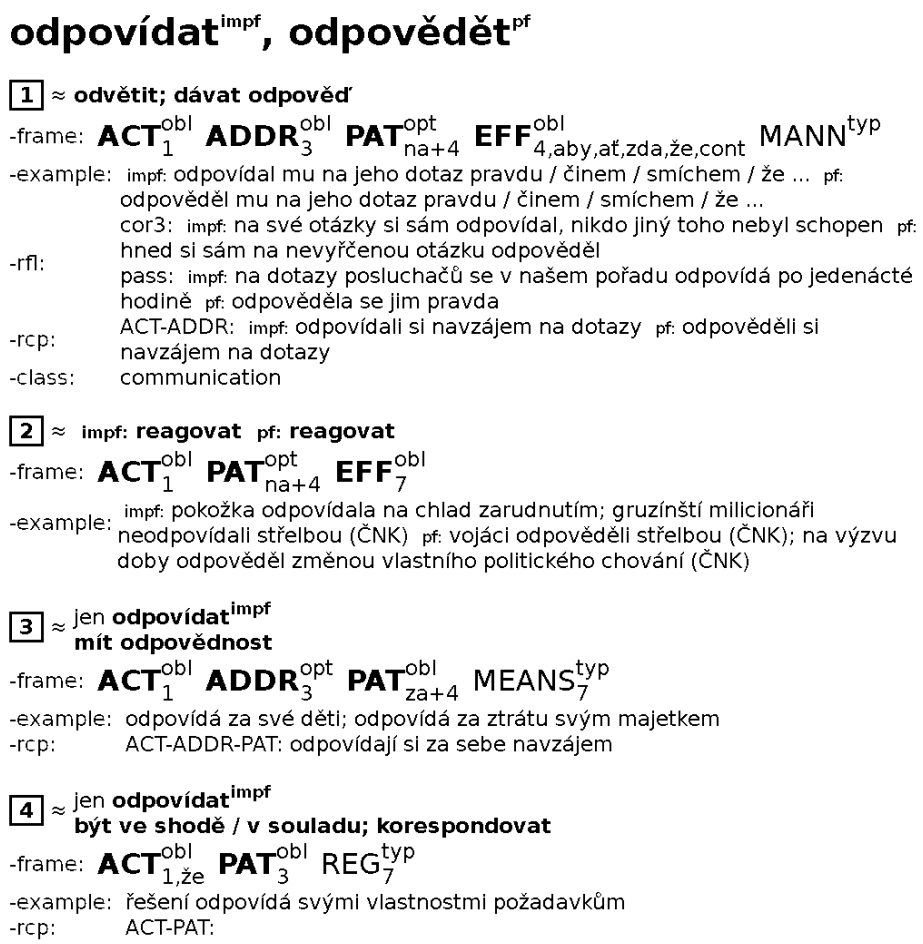
Frekvence. Ve slovníku je u každého valenčního rámce slovesa uvedena přibližná frekvence výskytu v textu. Ačkoli vzorek textu nebyl tak velký jako u VALLEXu, jde o čísla důležitá a unikátní, neboť z frekvence lexému se nedá usuzovat na frekvenci jednotlivých rámců a některý rámec nezařazený do VALLEXu může snadno být častější než jiný zařazený.

1.3 Odlišnosti

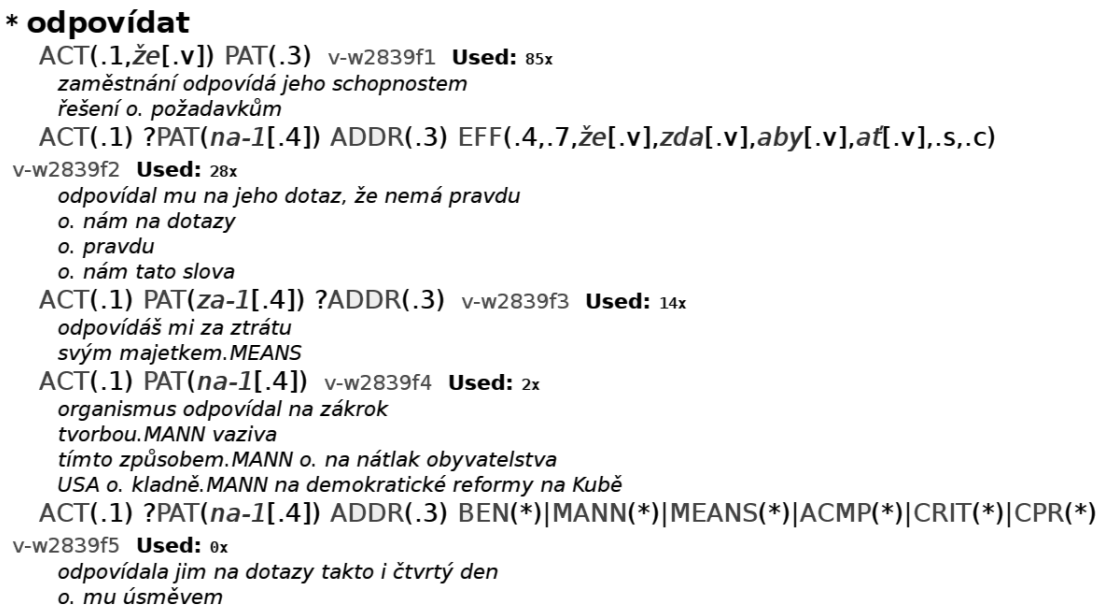
Na závěr kapitoly ještě projděme závažnější místa, v kterých se oba právě popsané slovníky liší.

1. Různé datové formáty. Oba slovníky jsou uloženy v XML, ale jejich struktura je zcela odlišná.
2. Jiný repertoár uvnitř rámce. Oba slovníky se trochu jinak vypořádávají se specifikací jednotlivých valenčních doplnění. Liší se mj. seznam spojek, které jsou použity k určení vedlejší věty ve valenci slovesa (PDT-VALLEX užívá navíc např. „jestli“), taktéž předložek (VALLEX užívá navíc např. „kolem“).
3. Různý přístup ke stejným rámcům. Slovníky vznikaly za jiných okolností a nemají tudíž sjednocenou metodologii. Tak se například stane, že věty „Dosáhl na něm slibu.“ a „Dosáhl svého.“ jsou v PDT-VALLEXu representovány dvěma rámci („dosáhnout na někom něčeho“ a frazém „dosáhnout svého“), kdežto ve VALLEXu jsou oba spojeny pod rámec jediný. Jiným příkladem je sloveso *napojit*, u něhož nastává homografie: sloveso nese jak význam „připojit, spojit“, tak také „dát napít“. V takovém případě je sloveso rozděleno na dva lexémy ve VALLEXu, ale ne v PDT-VALLEXu.

[2] Jsou to nejčastější česká slovesa vybraná podle jejich frekvence v Českém národním korupusu SYN2000: `http://ucnk.ff.cuni.cz`



Obrázek 1. Sloveso *odpovídat* se čtyřmi valenčními rámci, jak je zachyceno ve slovníku VALLEX. (Obrázek pochází z webového rozhraní: `http://ufal.mff.cuni.cz/vallex/2.5/data/html/generated/alphabet/index.html`.)



Obrázek 2. Sloveso *odpovídat* s pěti valenčními rámci, jak je zachyceno ve slovníku PDT-VALLEX. (Obrázek pochází z webového rozhraní ke slovníku: `http://ufal.mff.cuni.cz/pdt2.0/visual-data/pdt-vallex/0_vallex.html`.)

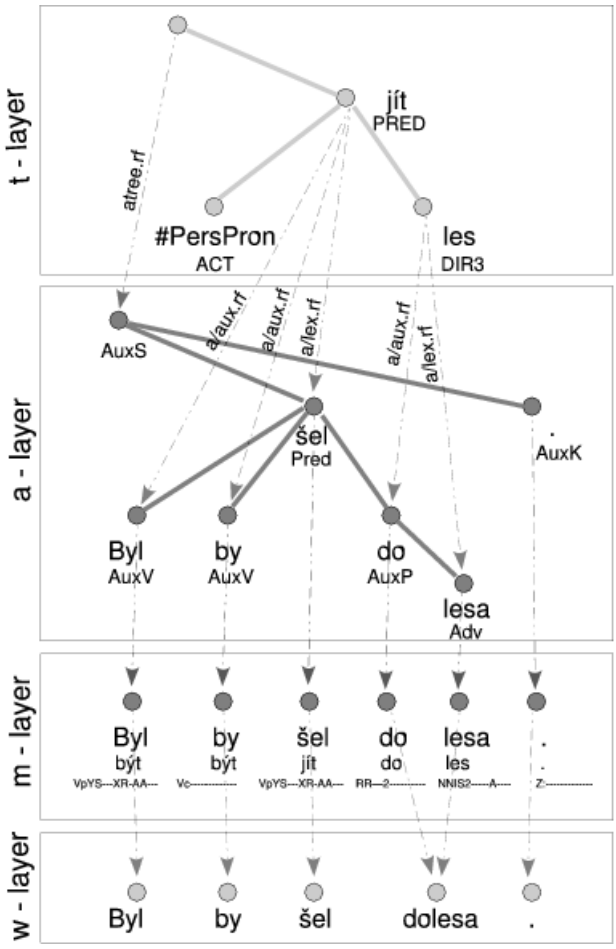
4. PDT-VALLEX zdaleka neobsahuje kompletní informaci pro jednotlivé lexémy. Ve slovníku jsou uvedeny jen ty rámce, které se vyskytly v datech, a pak některé další, které anotátoři chtěli uvést (ačkoli na ně v datech nenarazili). Nebylo cílem mít lexém zpracován v úplnosti. Navíc je sada informací u rámce o trochu chudší než u VALLEXu.
5. K VALLEXu neexistují žádná data, která by byla anotovaná odkazy na jeho rámce. Tato data musíme v našem projektu částečně vyrobit, abychom mohli přes data slévat některé odpovídající si rámce, nebo pak slévání evaluovat.

Body (1.) až (3.) pro nás jsou překážkou, kterou musíme při slévání překonat: (1.) je čistě technický problém, (2.) komplikuje automatické porovnání dvou shodných rámců (nebudou se jevit shodně, liší-li se metodologie jejich zachycení) a (3.) rovněž komplikuje automatickou proceduru a navíc vyžaduje lingvistické rozhodnutí, jak s takovými rámci zacházet při slévání. Naopak z bodů (4.) a (5.) plyne zisk po sloučení slovníků: výsledný slovník bude sjednocením obou současných a bude obsahovat odkazy do dat pro ty rámce z VALLEXu, pro které bude nalezen ekvivalent v PDT-VALLEXu.

2 Pražský závislostní korpus

Třetím lingvistickým datovým zdrojem, který využijeme, je Pražský závislostní korpus (Prague Dependency Treebank, PDT, [2]). Tento korpus českých vět je v souladu s FGP členěn do tří rovin a každá obsahuje novou vrstvu ruční anotace. Nejníže je *m-rovina* s anotací morfologickou a s lemmatisací jednotlivých slov. Nad ní leží *a-rovina* se stromovou strukturou zachycující větnou syntaxi, větné členy ap. Nejvýše je *t-rovina* s hloubkovou syntaxí, která obsahuje už pouze významová slova a přiřazuje jim značky jako actor, paciens atd. Příklad na obrázku 3 to ukáže nejlépe.

PDT obsahuje 2 miliony slov, z toho více než 830 000 je obohaceno anotací na všech třech rovinách. Ze všech jevů, které PDT popisuje, je pro nás podstatná hlavně valence. Každé slovo v PDT, které má valenci (tedy všechna slovesa a některá substantiva, adjektiva, adverbia), obsahuje odkaz k příslušnému valenčnímu rámci v PDT-VALLEXu (pokud takový rámec v okamžiku anotace ještě neexistoval, byl vytvořen).



Obrázek 3. Schematické znázornění všech tří rovin PDT a povrchové věty (*w-layer*) s překlepem: „Byl by šel dolesa.“ Uzly t-roviny nesou ještě mnoho další nezobrazené informace – kupříkladu uzel *jít* obsahuje odkaz `v-w1339f3` do PDT-VALLEXu.

2.1 Prague Markup Language

Dovolíme si krátkou odbočku k formátu, ve kterém jsou data PDT uchovávána a zpracovávána. Nazývá se Prague Markup Language (PML, [8]) a je to aplikace XML. Uchovává anotaci zvlášť po jednotlivých rovinách, takže informace o jedné větě jsou vždy ve čtyřech souborech (viz rámečky obrázku 3) a ty jsou mezi sebou prolinkované.

Také my chceme uchovávat informaci o provázání slovníků s daty a mezi sebou jako *stand-off* anotaci.[3] Proto využijeme tzv. s-soubory (které jsou taktéž instancí PML), které budou obdobou nové roviny anotace. Jejich obecný formát nám umožní odkazovat do dat i do slovníků najednou.

[3] Stand-off anotace je koncept, podle něhož se dodatečná informace uchovává odděleně od původních dat. Ta tudíž nejsou nijak modifikována a lze s nimi volitelně pracovat buď spolu s anotací, nebo bez ní.

3 Data

Pro náš projekt tedy využijeme dva valenční slovníky a korpus PDT, ve kterém je zanesena valenční informace z jednoho z nich.

VALLEX obsahuje asi 4800 lexémů, PDT-VALLEX 5500[4] (viz tabulka 1). Průnik, tedy lexémy obsažené (v nějaké formě) v obou slovnících, tvoří více než polovinu každého slovníku.

	lexémy	jedinečné	společné
VALLEX	4787	1618	3169
PDT-VALLEX	5510	2341	3169

Tabulka 1. Počty lexémů v obou slovnících

Dále spočítáme přibližnou obtížnost úlohy. Baseline je obvykle úspěšnost nějaké triviální metody počítaná proti správným (*gold standard*) datům. K této úloze ale neexistují ručně anotovaná data pro VALLEX a tak nemůžeme stanovit chybovost nějaké triviální metody. Spočítáme tedy alespoň pravděpodobnost, že při náhodném mapování lexému PDT-VALLEXu na rámce VALLEXu vybereme správný. V průměru má lexém VALLEXu 2,5 rámce. Pravděpodobnost, že zvolím ten správný (průměrovaná přes celý VALLEX) vychází 0,6. To dává naději, že chytřejší postup může být úspěšný.

4 Postup slévání

Slévání probíhá ve dvou hlavních fázích. V první se některé valenční rámce podaří namapovat na odpovídající rámce druhého slovníku automaticky. Ve druhé části bude potřeba slít zbylé rámce ručně. Na následujících příkladech si ukážeme, kdy můžeme rámce sloučit automaticky.

Například sloveso *kazit* má jen jeden rámec v každém slovníku. Ačkoli tyto rámce nejsou úplně stejné (ve VALLEXu je bohatší), jsou kompatibilní (rámec PDT-VALLEXu je podmnožinou) a je možné je sloučit.

Sloveso *kandidovat* má sice rámce dva, ale v obou slovnících. Navíc jsou prakticky shodné, odhlédneme-li od odlišností slovníků samotných. Pokud je přepíšeme ve sjednocené notaci, vypadá první rámec („Petr kandiduje na poslance (do Parlamentu ČR).“) postupně v obou slovnících takto:

$\mathsf{ACT}^{\mathsf{obl}}_{\mathsf{nom}}\ \mathsf{PAT}^{\mathsf{opt}}_{\mathsf{na+acc}}$
$\mathsf{ACT}^{\mathsf{obl}}_{\mathsf{nom}}\ \mathsf{PAT}^{\mathsf{opt}}_{\mathsf{na+acc,\ za+acc}}\ \mathsf{DIR3}^{\mathsf{typ}}$

a druhý („Kandidovali Petra na europoslance (do Bruselu).“) takto:

$\mathsf{ACT}^{\mathsf{obl}}_{\mathsf{nom}}\ \mathsf{PAT}^{\mathsf{obl}}_{\mathsf{acc}}\ \mathsf{EFF}^{\mathsf{opt}}_{\mathsf{na+acc}}$
$\mathsf{ACT}^{\mathsf{obl}}_{\mathsf{nom}}\ \mathsf{PAT}^{\mathsf{obl}}_{\mathsf{acc}}\ \mathsf{EFF}^{\mathsf{opt}}_{\mathsf{na+acc,\ za+acc}}\ \mathsf{DIR3}^{\mathsf{typ}}$.

Rámce jsou si dostatečně podobné, abychom je mohli sloučit (zvláště když ve VALLEXu žádný další rámec není a nemůže dojít k záměně).

Ke komplikovanějšímu řešení nás nutí sloveso *rozvinout*, které má v PDT-VALLEXu tři zcela shodné rámce ($\mathsf{ACT}_{\mathsf{nom}}\ \mathsf{PAT}_{\mathsf{acc}}$) a je potřeba je namapovat na dva rámce VALLEXu. Tyto dva rámce se liší nepovinným $\mathsf{MEANS}_{\mathsf{inst}}$ – a toho zkusíme využít. Vyhledáme sloveso *rozvinout* v PDT, kde je u každého odkaz na jeden ze tří rámců PDT-VALLEXu. Pokud u některého z nich bude na slovese záviset také větný člen MEANS v instrumentálu, bude to náš kandidát.

Úspěšnost celého automatického postupu budeme testovat na ručně namapovaných rámcích a na datech, kterým vedle valenčních rámců PDT-VALLEXu přiřadíme ručně rámce z VALLEXu.

Výsledkem celého procesu pak například pro naše ukázkové heslo *odpovídat* budou odkazy

- od prvního rámce v PDT-VALLEXu (obr. 2) na 4 ve VALLEXu (obr. 1),
- od druhého rámce na 1,
- od třetího rámce na 3,
- od čtvrtého rámce na 2 a
- od pátého rámce opět na 1.

Celá procedura však má být použitelná i po vydání rozšířených versí slovníku. Proto musí umožňovat na závěr použít automatickou proceduru z první fáze a data z ruční druhé (ruční) fáze a provést celé slití slovníku automaticky.

Je tedy potřeba heuristika, která při novém slévání pozměněného rámce vezme v úvahu minulý výsledek a rozhodne, zda je změna dostatečně malá, aby mohlo sloučení proběhnout totožně. Pro nezměněné rámce bude výsledek stejný. Na ruční práci tedy zbydou pouze nové a zároveň problémové rámce a rámce příliš odlišné od předchozí verse slovníku. Pokud tento postup použijeme na původní nezměněné slovníky, musí být výsledný slovník totožný s ručně upraveným; to poslouží jako kontrola.

5 Závěr

Představili jsme projekt, který má propojením dvou valenčních slovníků získat nový kvalitní lexikografický zdroj. Z odhadované pravděpodobnost náhodného mapování lze usuzovat na použitelnou úspěšnost automatické procedury.

[4] To je počet pouze slovesných lexémů, dalších 4528 tvoří substantivní, adjektivní a adverbiální lexémy.

Poděkování

Tato práce je podporována granty Grantové agentury Univerzity Karlovy č. 4200/2009 a Grantové agentury Akademie věd ČR č. 1ET100300517 a projektem MŠMT ČR LC536.

Reference

1. Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová-Řezníčková, V., Pajas, P. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. Proceedings of The Second Workshop on Treebanks and Linguistic Theories (2003) 57–68
2. Hajič, J. et al.: Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia, PA, USA (2006) URL: `http://ufal.mff.cuni.cz/pdt2.0`
3. Lopatková, M., Žabokrtský, Z., Kettnerová, V.: Valenční slovník českých sloves. Karolinum, Praha (2008) 382 str. URL: `http://ufal.mff.cuni.cz/vallex/2.5`
4. Panevová, J.: On Verbal Frames in Functional Generative Description, Part I. The Prague Bulletin of Mathematical Linguistics **22** (1974) 3–40
5. Panevová, J.: On Verbal Frames in Functional Generative Description, Part II. The Prague Bulletin of Mathematical Linguistics **23** (1975) 17–52
6. Panevová, J.: Valency Frames and the Meaning of the Sentence. The Prague School of Structural and Functional Linguistics, Amsterdam, Philadelphia (1994) 223–243
7. Sgall, P.: Generativní popis jazyka a česká deklinace. Praha, Academia (1967)
8. Pajas, P., Štěpánek, J.: A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague Dependency Treebank 2.0. ÚFAL Technical Report **29**, Praha: MFF UK (2005)