VALENČNÍ SLOVNÍK STOKRÁT JINAK: CO JE POD POVRCHEM?

(VALENCY LEXICON: WHAT CAN BE FOUND UNDER THE SURFACE?)

Markéta Straňáková-Lopatková, Zdeněk Žabokrtský
Centrum komputační lingvistiky, MFF UK Praha

Pro automatické zpracování jazyka (NLP) stejně jako pro teoretickou lingvistiku potřebujeme obsáhlý valenční slovník. Vytváření takového slovníku českých sloves vyžaduje systematické zachycení všech syntaktických údajů potřebných pro analýzu češtiny. Budovaný slovník proto obsahuje různé typy informací – kromě jevů patřících k popisu hloubkové struktury (zejména hloubkových rolí jednotlivých doplnění ('funktorů')) i údaje morfematické (realizace jednotlivých doplnění). Důraz je kladen na maximální systematičnost a konzistenci zpracovávaných jevů.

Pro češtinu již existují valenční slovníky, nicméně buď je jejich rozsah omezený a forma neumožňuje automatické zpracování (Slovesa pro praxi), nebo nejsou dostatečně spolehlivé díky automatickému zpracování (Skoumalová), případně vůbec neobsahují podkladovou strukturu (Horák). To vedlo k projektu vytvoření valenčního slovníku, který by čerpal z existujících slovníků a splňoval všechny uvedené požadavky.

Vycházíme z klasifikace slovesných doplnění, která je důkladně zpracována ve Funkčním generativním popisu češtiny (např. Panevová (1974-75), (1994)). V tomto pojetí se valenční rámec skládá z vnitřních doplnění, **aktantů (obligatorních i fakultativních)** a z **obligatorních adverbiálních (volných) doplnění**.

Vzhledem ke snaze o co nejbohatší slovníkové informace pro NLP zavádíme jemnější členění volných doplnění – zavádíme navíc třídu tzv. **kvazivalenčních doplnění**, která z teoretického hlediska patří k fakultativním volným doplněním, nicméně jde o "obvyklá doplnění" charakteristická pro konkrétní slovesa, která mohou specifikovat jeho význam, příp. jeho posunuté či přenesené užití (např. volné doplnění směru u slovesa *jet*, doplnění prostředku (Means) u slovesa *hrát – hrát na kytaru*). Kromě kvazivalenčních doplnění jsou ve slovníku uváděna též **typická volná doplnění** (na základě příkladů především ze Slovníku spisovného jazyka českého (SSJČ), např. Benefaktor slovesa *čekat – čekat někomu (s dluhem)*).

Přehled údajů zachycených ve slovníku

• Ke každému slovesu je dán seznam valenčních rámců (každé sloveso má alespoň jeden rámec, může jich však mít víc, podle počtu jeho "významů"; přitom se bere v úvahu primární, přenesené i idiomatické užití slovesa).

• Každý slovesný rámec obsahuje především výčet jednotlivých doplnění; kromě toho jsou specifikovány i některé další údaje, např. odkaz na jednotku sémantické databáze EuroWordNet (EWN, viz níž), sémantická třída (verba dicendi, ...), příklad užití, lemma čistě vidového protějšku.

• Jednotlivá slovesná doplnění jsou charakterizována příslušným funktorem (jménem aktantu nebo volného doplnění), typem doplnění (obligatorní, fakultativní, kvazivalenční, typické) a jeho morfematickým vyjádřením. Dále je uváděna možnost recipročního užití daného doplnění.

Vznik slovníku a jeho zdroje

Jako první testovací vzorek byla zpracována skupina 160 nejčastějších českých sloves (kromě slovesa *být* a modálních sloves). Nyní postupně zpracováváme další "dávky" po 100 slovesech (říjen 2001: 200 sloves hotovo, 300 sloves rozpracováno). Kritériem výběru je četnost lemmatu v Pražském závislostním korpusu (PDT).

Postup zpracování každé dávky se skládá ze dvou kroků: z automatického předzpracování dat a z ruční anotace. V prvním kroku jsou vybraná slovesa nalezena v "brněnském valenčním slovníku" (obsahuje morfematické údaje o možných slovesných doplněních, byl vytvořen na základě SSJČ, viz Horák (1998)) a do takto získaných rámců jsou předvyplněny funktory. Výsledek je převeden do formátu XML. Do téhož souboru jsou přidány informace z dalších zdrojů (EWN, výsledky z tektogramatické anotace PDT (Hajičová et al (2000), seznamy teoreticky zpracovaných skupin sloves). Při ruční anotaci jde především o následující operace: slučování a rozdělování rámců, doplňování prvků rámce, opravy funktorů a povrchových forem, přidání příkladu užití rámce.

Zkoumáme možnost napojení jednotlivých slovesných rámců na budovanou sémantickou síť EuroWordNet (Pala a Ševeček (1999)). EuroWordNet je databáze, která na úrovni lexikálních jednotek propojuje angličtinu a sedm dalších evropských jazyků. Pro češtinu je zatím zpracováno cca 3000 sloves.

Možnosti využití slovníku

V současné době slovník napomáhá při tektogramatické anotaci Pražského závislostního korpusu, zejména pro udržení konsistence při zpracovávání sloves. Slouží také jako zpětná vazba pro ověřování kvality EuroWordNetu. Předpokládáme, že bude využit pro lexikální disambiguaci a pro syntaktickou analýzu – takto koncipovaný valenční slovník umožní podstatným způsobem zkvalitnit automatickou analýzu češtiny (pokud je taková analýza vůbec možná bez slovníku tohoto typu). V budoucnosti (díky napojení na další jazyky prostřednictvím EuroWordNetu) by se měl stát významnou pomůckou při strojovém překladu.

Literatura

HAJIČOVÁ, E., PANEVOVÁ, J., SGALL, P. (2000) *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank.* ÚFAL/CKL Technical Report TR-2000-09.

HORÁK, A. (1998) Verb valency and semantic classification of verbs. In: *Proceedings of TSD'98*, Masaryk University Press, Brno, s.61-66.

PALA, K., ŠEVEČEK, P. (1999) *Final Report, Final CD ROM on EWN1,2LE4-8328*, Amsterdam.

PANEVOVÁ, J. (1974, 1975) On Verbal Frames in Functional Generative Description. Part I, *PBML* 22, pp.3-40, Part II, *PBML* 23, s. 17-52.

PANEVOVÁ, J. (1994) Valency Frames and the Meaning of the Sentence. In: *The Prague School of Structural and Functional Linguistics* (ed. P. A. Luelsdorff), Benjamins Publ. Comp. Amsterdam, Philadelphia, s. 223-243.

H. SKOUMALOVÁ (2001) *Czech syntactic lexicon.* PhD thesis, Charles University, Faculty of Arts, Prague.

SVOZILOVÁ, N., PROUZOVÁ, H., JIRSOVÁ, A. (1997) *Slovesa pro praxi*, Academia, Praha.