Valence a její formální popis. Vybrané aspekty budování slovníku *VALLEX*[*]

Markéta Lopatková

ÚFAL MFF UK
Malostranské nám. 25, Praha
lopatkova@ufal.mff.cuni.cz

Abstrakt. Valencí se rozumí schopnost slovesa (příp. slova jiného slovního druhu) vázat na sebe určitý počet jiných, syntakticky závislých jazykových jednotek. Valenční informace se tedy vztahuje k jednotlivým lexémům, a jako takovou je nutno popsat ji pro jednotlivé lexémy ve formě slovníku. Bez valenčních slovníků se neobejdou komplexní aplikace pro zpracování přirozeného jazyka, které jsou založeny na explicitním popisu jazykových jevů; zároveň jsou takové slovníky nepostradatelné při vytváření jazykových dat, na nichž jsou založeny nástroje využívající strojového učení.

V tomto textu shrnujeme výsledky dosažené při vytváření lexikální databáze českých sloves. Práce se soustřeďuje na tři základní okruhy. Prvním okruhem je formální zachycení valenčních vlastností českých sloves ve valenčním slovníku. Je zde představena logická stavba bohatě strukturovaných slovníkových dat. Druhým okruhem, kterému se práce věnuje, jsou nové teoretické aspekty, které přináší zpracování rozsáhlého jazykového materiálu – je to především koncept kvazivalenčních doplnění a adekvátní zpracování slovesných alternací. Třetí okruh tvoří problematika formálního modelování přirozeného jazyka. Je zde představen nový formální model závislostní syntaxe založený na originálním konceptu restartovacích automatů.

Hlavním aplikovaným výstupem této práce je *Valenční slovník českých sloves VALLEX*, rozsáhlý a kvalitní veřejně dostupný slovník, který obsahuje významové a valenční charakteristiky nejčastějších českých sloves. Při navrhování jeho koncepce byl kladen důraz na možnost všestranného využití pro člověka jako uživatele jazyka i pro aplikační účely při automatickém zpracování češtiny.
Klíčová slova: valence sloves, VALLEX, korpus, valenční syntax, redukční analýza, restartovací automaty

Úvodní poznámky

Jazykový systém se obvykle dělí na dvě základní složky, gramatiku a lexikon. *Gramatika* popisuje v zásadě obecné zákonitosti přirozeného jazyka pomocí pravidel,

[*]Tato práce vznikla v rámci Výzkumného záměru Ministerstva vnitra, mládeže a tělovýchovy ČR č. MSM0021620838 a dále za částečné podpory Grantové agentury Akademie věd ČR, program Informační společnost č. 1ET100300517.

jež lze uplatňovat na celé třídy slov (specifikované obvykle morfologicky či syntakticky, někdy do jisté míry i sémanticky). Naproti tomu *lexikon* obvykle zachycuje ty charakteristiky jazykového systému, které jsou vázány na jednotlivé lexikální jednotky jazyka. Tyto charakteristiky lze opět rozdělit podle úrovně popisu jazyka na charakteristiky morfologické, zachycované v morfologických slovnících, a na charakteristiky popisující kombinatorický potenciál jednotlivých (významových) slov, tedy rysy syntaktické a syntakticko-sémantické; tento druhý typ charakteristik lexikálních jednotek je popisován ve valenčních slovnících.

Jednotlivé teorie popisující jazykový systém se různí v tom, jaká část informace, potřebné pro popis přirozeného jazyka, je zachycena pomocí obecných pravidel (tedy v gramatické komponentě jazykového systému) a jakou její část je vhodné uchovávat v komponentě slovníkové. V současném vývoji formální a aplikované lingvistiky lze vysledovat příklon k rozsáhlým a bohatým slovníkovým informacím (v teoretických studiích se tato tendence odráží v pojmech *lexikalizovaná gramatika*, viz [26], či *valenční syntax*, viz [78, 79]).

Následující text přístupnou formou shrnuje základní problémy, které vznikají při budování valenčního slovníku, a jejich řešení přijatá při vytváření *Valenčního slovníku českých sloves* (dále *VALLEX* jako *VALency LEXicon*). Jde o téma vysoce aktuální, neboť s příklonem ke zkoumání vyšších rovin jazyka, hloubkové syntaxe a sémantiky, se pozornost lingvistů soustřeďuje na teoretický popis valence (a to nejen jazykových jevů centrálních, které jsou zkoumány již od poloviny minulého století, ale také jevů na hranici centra a periferie i jevů čistě periferních) a na zachycení valenčního chování jednotlivých lexémů ve slovnících.

Problematika a popis valence není pouze záležitostí teoretické lingvistiky – bez valenčních slovníků se neobejdou ani pokročilé aplikace pro zpracování přirozeného jazyka (dále NLP, Natural Language Processing), které jsou založeny na explicitním popisu jazykových jevů (často označované jako ‚rule-based' přístupy). Zároveň jsou takové slovníky nepostradatelné při vytváření jazykových dat, na nichž jsou založeny nástroje využívající strojového učení (‚data-driven' přístupy).

Impulsem autorky k návrhu a řešení projektu zaměřeného na vytváření lexikální databáze českých sloves *VALLEX*, popisované v tomto textu, byla neexistence rozsáhlého a kvalitního veřejně dostupného slovníku, který by obsahoval významové a valenční charakteristiky českých sloves a přitom byl široce uplatnitelný v NLP aplikacích.

Tento text je členěn do šesti oddílů. V oddíle 1 je nejprve přiblížen pojem (slovesné) valence.

Oddíl 2 představuje základní strategie, jak získat valenční slovník. Je to především možnost převést existující tištěný slovník do elektronické podoby a obohatit jej o významovou reprezentaci. Alternativní možností je automatická extrakce slovníku ze syntakticky či tektogramaticky anotovaného korpusu. Dále je zde ilustrován vztah mezi složitostí valenčního chování sloves a jejich frekvencí v korpusu. Oddíl 2 uzavírá stručný popis první zveřejněné verze slovníku *VALLEX 1.0* (a jeho kvalitativně i kvalitativně rozšířené verze *1.5*).

Oddíl 3 je věnován teoretickým aspektům valence a jejich promítání do koncepce slovníku. Zabývá se především originálním konceptem kvazivalenčních doplnění, dále

pak problematikou rozlišení jednotlivých významů slovesa. Dalšími tématy tohoto oddílu jsou zpracování sloves s podobnými sémantickými vlastnosti a zejména pak návrh alternačního modelu slovníku.

Oddíl 4 přibližuje současnou verzi slovníku *VALLEX 2.0*.

Oddíl 5 je věnován formálnímu modelování přirozeného jazyka, kde valence slouží jako základní syntaktická informace určující závislostní strukturu věty. Je zde představena metoda redukční analýzy pro závislostní syntax. Dále je zde neformálně představen koncept restartovacího automatu, který vhodným způsobem modeluje valenční syntax a nelokální chování jazyků s volným slovosledem. Redukční systém založený na konceptu restartovacího automatu představuje nový formální rámec pro modelování Funkčního generativního popisu, teoretické koncepce popisu přirozeného jazyka, která tvoří teoretický základ této práce.

Text uzavírá oddíl 6 shrnující dosavadní využití slovníku *VALLEX* v NLP aplikacích, zejména při budování jiných elektronických slovníků. Jsou zde též zmíněny další směry rozvíjení slovníku, především postupné zpřesňování kritérií pro rozlišování jednotlivých významů slov a dále promítnutí alternačního modelu slovníku do dat.

1 Co je to valence?

Termínu valence známého z oblasti chemie, kde označuje mocenství atomu, v lingvistickém kontextu poprvé užil v polovině minulého století francouzský syntaktik Lucien Tesnière. Metaforicky tímto termínem označil schopnost slovesa (podobnou vlastnosti atomů) vázat k sobě určitý počet jazykových elementů.

Pod pojmem valence se rozumí „počet a povaha míst (argumentů), které na sebe dané sloveso (popř. slovo jiného slovního druhu) váže“ (*Encyklopedický slovník češtiny*, viz [28]). Je to tedy schopnost slova vázat na sebe určitý počet jiných, syntakticky závislých jazykových jednotek. Tato schopnost se primárně týká významové roviny jazyka, tedy hloubkové větné stavby. Valenční pozice jsou naplňovány valenčními doplněními, jako je aktor (konatel či nositel děje, dále ACT), patient (zasažený objekt, PAT), adresát (ADDR), původ (ORIG) a výsledek děje (EFF), označovanými obvykle jako aktanty, ale i volnými doplněními vyjadřujícími okolnosti děje, jako jsou čas, místo, směr, způsob děje apod.

Soubor valenčních pozic, které charakterizují slovo v jednom z jeho významů, je označován jako valenční rámec daného slova v daném významu.[1]

Jednotlivé valenční pozice jsou různě významově těsné. Neobsazením některých pozic dochází k porušení významové úplnosti, což může vést i k porušení gramatické správnosti věty, srov. např. nepřijatelné věty **Petr dává, *Marie nenávidí, *Jan se choval*. Doplnění obsazující takové valenční pozice se nazývají obligatorní (na významové rovině popisu). Zvláštní pozornost je v teorii valence věnována případům, kdy obligatorní valenční pozice zůstávají v povrchové podobě věty neobsazeny a posluchač / čtenář si příslušné pozice zaplňuje z kontextu promluvy / textu, např. ve spojeních *obvykle nakupuje v supermarketu* či *děti přišly* není vyjádřen patient,

[1] Vztah mezi jednotlivými významy slova a jeho valenčními rámci je obecně složitější, více viz [47].

resp. směr – tyto informace by měly být posluchači / čtenáři zřejmé z kontextu promluvy / textu.

Jiné valenční pozice jsou nepovinné, fakultativní – jsou sice přítomny ve valenčním rámci a mohou být přítomny ve významové reprezentaci věty, jejich neobsazením však nevzniká významově ani gramaticky narušená věta, např. *Petr se pevně držel (zábradlí)*, *Eva se najedla (ovoce)*, *Dívka píše (mamince) dopis*. Ostatní pozice s velmi volným vztahem ke slovesu se často označují jako nevalenční či volné, např. *Jana se procházela (po lese)*, *Petr se budil (časně)*, *Eva si četla (pro své potěšení)*, i když v teoretickém popisu jsou do valence slovesa (v širším smyslu) zahrnuty.

V povrchové realizaci věty jednotlivé aktanty vlivem slovesa obvykle nabývají určité formy – jejich morfématická podoba je určována požadavky řídícího slovesa (zpravidla jeho rekcí). Např. aktor bývá v aktivní větě typicky vyjádřen nominativem, zatímco patient se obvykle realizuje akuzativem (např. $Petr_{ACT}$ *ztratil* $botu_{PAT}$), jiná slovesa vyžadují aktora v dativu (např. $Petrovi_{ACT}$ *se ve škole líbí*), opět jiná slovesa mají patient v dativu (např. $rodiče_{ACT}$ *bránili jejich* $štěstí_{PAT}$) či ve formě předložkové skupiny (např. *doufali ve* $vítězství_{PAT}$). Naproti tomu forma valenčních volných doplnění bývá dána významem těchto doplnění (např. *děti přišly domů / do školy / na hřiště*) a nebývá řízena slovesem.

Valence je teoreticky zkoumána zhruba od poloviny 20. století, jmenujme zde již zmíněného L. Tesnièra [88]. Důraz na zkoumání významové složky valence pak přinesly především studie Ch. Fillmora [11, 12]. V českém prostředí se k významným pracím věnovaným valenci řadí zejména studie F. Daneše, například [7, 9], práce J. Panevové, zejména [58, 59, 60, 62], P. Karlíka [27], a práce P. Sgalla [78, 79]. Další odkazy lze nalézt v práci [47].

Sloveso je tradičně považováno za organizační centrum věty, neboť svými požadavky na počet a povahu syntakticky závislých jazykových jednotek, tedy svými valenčními požadavky, vytváří její (relační) strukturu (*Mluvnice češtiny 3* [10]). Proto se teoretické zkoumání valence zaměřuje primárně na slovesa, valence dalších autosémantických slovních druhů (substantiv, adjektiv a adverbií) je obvykle chápána jako sekundární.

Je zřejmé, že valenční vlastnosti slov jsou velmi rozmanité. Nelze je odvodit obecnými pravidly, je třeba je popsat pro jednotlivé lexikální položky, tedy v podobě valenčního slovníku, který obsahuje popis valence jednoho slova po druhém, v každém z jejich významů. Verbocentrický přístup v teoretické syntaxi se promítá i do úsilí budovat primárně valenční slovníky pro slovesa; při popisu valence substantiv a adjektiv se potom typicky využívá strukturní podobnost (deverbativních) jednotek se slovesy.

Dodejme, že budování valenčních slovníků jak v tištěné, tak v elektronické podobě stojí v současné době v centru pozornosti lexikografie české i lexikografie dalších desítek jazyků; odkazy na nejdůležitější slovníky jsou uvedeny v práci [47], přehled a anotaci nejvýznamnějších projektů zabývajících se valencí lze nalézt v [90].

2 Postupné budování slovníku *VALLEX*

Rozvoj počítačové lingvistiky a zájem o aplikační úkoly přináší potřebu lingvistických dat s různými typy lingvistických informací (morfologie, povrchová či hloubková syn-

tax, rozlišení významů apod.). Zejména díky Pražskému závislostnímu korpusu (PDT, Prague Dependency Treebank, viz [15]) se čeština řadí k jazykům s nejbohatšími datovými zdroji.

Valenční slovník představuje další významný datový zdroj, který poskytuje důležité informace pro mnohé úkoly NLP. Valence hraje základní roli v komplexních aplikacích pracujících s významem, jmenujme zde např. automatický překlad či sumarizaci, pro něž je syntaktická analýza a automatické rozlišování významů jednotlivých autosémantických slov jedním ze základních předpokladů.

Převod tištěného slovníku a jeho obohacení o významovou reprezentaci

Základní možností, jak získat valenční slovník, je převedení tištěného slovníku do elektronické podoby a následná extrakce valenční informace. Takovýmto způsobem vznikl slovník *BRIEF* [56], který vychází především ze *Slovníku spisovného jazyka českého*, dále SSJČ [19]. Textová data byla automaticky zpracována a ze slovníkových příkladů byly extrahovány možné povrchové kombinace valenčních doplnění (přitom však byly sloučeny jednotlivé významy sloves). Takto získaný slovník se omezuje na povrchové vyjádření valence a neřeší obligatornost jednotlivých valenčních doplnění. Jako největší nedostatek se ovšem jeví ztráta informace o jednotlivých slovesných významech. Formální zápis rámců, popsaný např. v článku [24], umožňuje využít takto získaný slovník pro NLP.

Přirozeným pokusem, jak slovník *BRIEF* obohatit o chybějící valenční informaci, byla automatická delimitace významů a doplnění významové charakteristiky ve formě funktorů pro jednotlivé valenční pozice. O takové automatické obohacení slovníku se pokusila H. Skoumalová [84]. Pomocí algoritmů, které autorka navrhla, byla posléze zpracována testovací sada sloves, označovaná též jako *Vallex-00*. Tato sada obsahovala 178 nejčastějších českých sloves (s výjimkou slovesa *být* a modálních sloves), která byla nejprve zpracována automaticky a poté dále rozsáhle upravována manuálně. Tato fáze budování slovníku je podrobněji popsána v článku [85]. Na zkoumaném vzorku sloves se však ukázalo, že nutné ruční úpravy[2] jsou takového rozsahu, že je ekonomičtější a pro anotátory jednodušší budovat slovník ručně.

Valenční informace a syntakticky anotovaný korpus

Další možnost, jak získat valenční slovník, spočívá ve využití existujících syntakticky anotovaných korpusů. V literatuře je popsáno několik technik pro automatické získávání (alespoň povrchové) valenční informace (‚subcategorization frames') z anotovaných korpusů. Také pro češtinu byly tyto techniky vyzkoušeny – jmenujme zde především práce [93, 74], které referují o vytváření povrchových valenčních rámců z analytické roviny PDT [14].

Takto získané rámce ovšem neobsahují významovou charakteristiku jednotlivých valenčních pozic ani neřeší jejich obligatornost. Opět největším nedostatkem je

[2] Šlo zejména o úpravy týkající se vymezování slovesných významů – přibližně 350 automaticky navržených rámců bylo anotátory upraveno na více než 460 rámců, které zhruba odpovídají jednotlivým významům sloves.

skutečnost, že tyto rámce nekorespondují s jednotlivými významy sloves.

Valenční informace a tektogramaticky anotovaný korpus

Jako třetí možnost, jak získat valenční slovník, se jeví jeho přímočará extrakce z tektogramaticky anotovaného korpusu. Základní idea je jednoduchá – budeme-li mít korpus, který je anotován školenými anotátory-lingvisty na tektogramatické rovině popisující jazykový význam věty, budeme mít u každého slovesa (a též substantiva či adjektiva s valenčními požadavky) zachycenu též informaci o jeho valenčním chování, tedy informaci o počtu a typu valenčních pozic a sekundárně i o jejich obligatornosti. Stačí potom shromáždit tuto informaci ve valenčním slovníku.

Při rozsáhlých manuálních anotacích tektogramatické roviny PDT [83, 15] se však ukázalo, že anotátoři – byť zkušení lingvisté – ke své práci nutně potřebují valenční slovník, neboť se během komplexní tektogramatické anotace nemohou soustřeďovat na problematiku jednotlivých jevů. Bylo proto rozhodnuto o manuálním budování valenčního slovníku s co největší technickou podporou (vhodné anotační prostředí, sebrané elektronické zdroje, vyhledávání podle nejrůznějších kritérií, kontroly konzistence apod.).

V první fázi byly shromážděny valenční rámce z již zmíněné sady sloves tvořících *Vallex-00*, viz [85], a pomocné seznamy, které používali anotátoři při anotaci PDT (prosinec 2001). Takto získaný materiál byl důkladně zpracován – především byly identifikovány odpovídající si rámce a byla vytvořena pomocná kritéria pro rozlišování významů. Takto zpracovaná slovesa (celkem 331 sloveso) se stala jádrem pro budování elektronického valenčního slovníku.

Vzhledem k nutnosti využívat valenční slovník pro anotaci PDT pokračovalo jeho budování ve dvou větvích. Obě tyto větve, včetně teoretického pozadí, jsou přiblíženy v článku [39], zde podáváme jen jejich velmi krátkou charakteristiku.

PDT-VALLEX. Slovník *PDT-VALLEX* zachycuje valenční rámce sloves, která se vyskytla při anotaci PDT, a to až na výjimky pouze v těch významech, v jakých se daná slovesa v PDT objevila.[3] Vznikal postupně jako anotační pomůcka, po ukončení anotací byl podroben rozsáhlé kontrole konzistence a jeho úpravy byly zpětně promítnuty do dat PDT. Zde se touto větví slovníku už dále zabývat nebudeme, více viz [16, 89].

VALLEX. Slovník *VALLEX* zpracovává celé slovesné lexémy, jeho cílem je popsat valenční chování slovesa ve všech jeho významech.[4] Důraz je kladen na teoreticky adekvátní popis valenčního chování, na relativní úplnost a na konzistenci zpracování. Kromě valenčního rámce obsahuje ke každému významu slovesa též další syntaktické informace, které souvisejí s povrchovými projevy valence (např. reciprocita, reflexivita či gramatická kontrola), a též některé informace syntakticko-sémantické (zejména syntakticko-sémantická třída).
Současná podoba a rozsah slovníku *VALLEX* jsou podrobně popsány v úvodu k tištěnému slovníku [47].

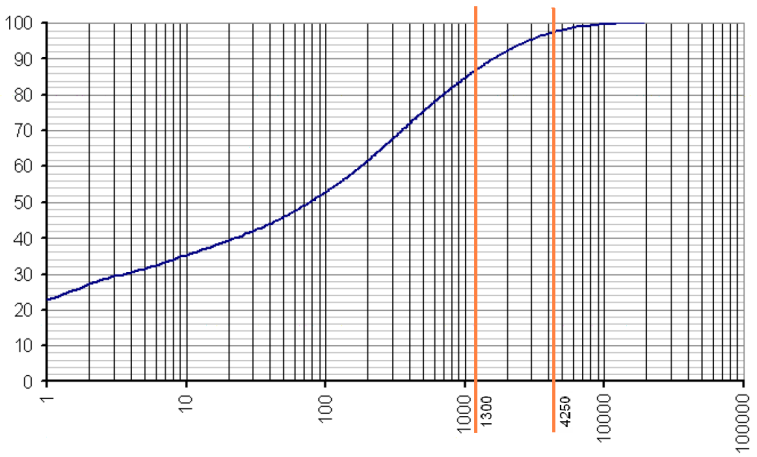
[3] *PDT-VALLEX* navíc obsahuje i valenční rámce některých deverbativních substantiv a adjektiv.
[4] *VALLEX* se soustřeďuje na primární i přenesené významy; popis idiomů zde není vyčerpávající.

V současné době se připravuje projekt, který obě větve slovníku, *PDT-VALLEX* i *VALLEX*, vzájemně (polo)automaticky propojí, čímž vznikne cenný jazykový zdroj popisující valenční chování (zejména) sloves, který bude široce provázaný s korpusovou anotací.

Poznamenejme, že v téže době a s týmiž problémy vznikal přibližně stejně velký valenční slovník pro angličtinu, *PropBank Lexicon* [32], spojený s korpusem Proposition Bank (PropBank, viz [57]), který je založen na anotaci tzv. propozic a jejich argumentové struktury v části korpusu Penn Treebank [52].

Pokrytí slovesných výskytů v korpusu a složitost valenčního chování sloves

Manuální vytváření rozsáhlého valenčního slovníku je časově náročné, zajišťuje však potřebnou konzistenci a adekvátnost popisu sloves. Analýza českých textů ovšem ukazuje, že pokrytí textů pro nejčastější česká slovesa splňuje tzv. Zipfův zákon [94], tedy že jednotlivá slovesa se vyskytují v určitém statistickém rozdělení – zhruba řečeno, frekvence slovesa v korpusu je nepřímo úměrná jeho relativnímu pořadí (při frekvenčním uspořádání), viz obrázek 1.

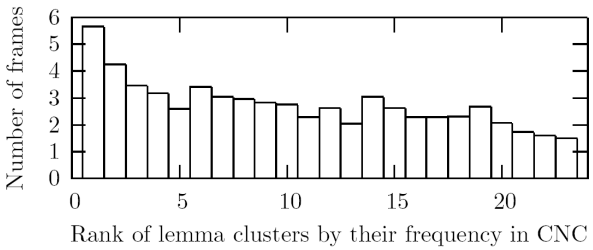


Obrázek 1: Pokrytí subkorpusu ČNK slovesnými lemmaty, převzato z dizertační práce [90]. Graf zachycuje na horizontální ose (s logaritmickým měřítkem) počet slovesných lemmat (bez případného reflexivního morfému *se/si*), vertikální osa pak udává kumulativní procentuální pokrytí korpusu.[5] Z tohoto obrázku je zřejmé, že např. valenční slovník obsahující 1 300 nejčastějších českých sloves pokryje zhruba 85% textů v subkorpusu Českého národního korpusu (ČNK, SYN2000)[6], což přibližně odpovídá slovníku *VALLEX 1.0* s přidaným slovesem *být* a s modálními slovesy (jednotlivé verze slovníku jsou popsány níže). Valenční slovník popisující 4 250 českých sloves potom pokrývá více než 96% výskytů slovesných lemmat v tomto subkorpusu (této velikosti odpovídá slovník *VALLEX, verze 2*, viz oddíl 4).

[5] V tomto orientačním grafu počítáme všechny výskyty sloves, včetně těch výskytů, kde se jedná o pomocná slovesa.

[6] `http://ucnk.ff.cuni.cz/`

Zkoumáme-li dále složitost valenčního chování sloves, zjišťujeme další významnou charakteristiku – obecně čím má sloveso vyšší frekvenci v citovaném korpusu, tím více má valenčních rámců a vykazuje tedy tím složitější valenční chování, viz graf na obrázku 2.



Obrázek 2: Počet valenčních rámců pro lexémy ve slovníku *VALLEX 1.0* a jejich frekvence v ČNK, převzato z článku [3].

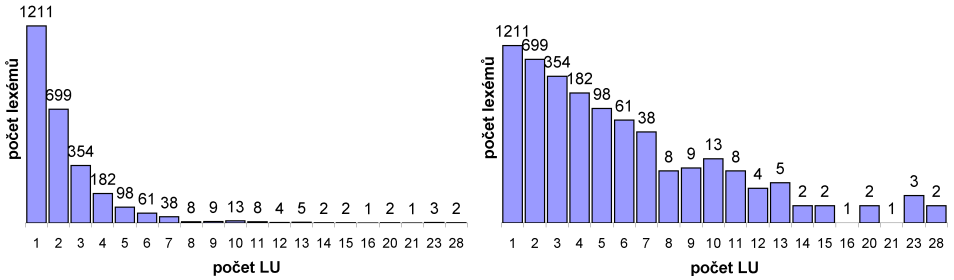
Jednotlivé lexémy sdružující vidové protějšky sloves (v grafu na obrázku 2 užit termín ‚lemma clusters')[7] byly uspořádány podle klesající frekvence a rozděleny do skupin po čtyřiceti lexémech. Sloupečky zobrazují průměrný počet valenčních rámců pro tyto skupiny lexémů (‚number of frames').[8]

Podívejme se tedy ještě na další jev spojený s valencí sloves, a to na distribuci valenčních rámců, které odpovídají jednotlivým lexikálním jednotkám (LU, lexical units, viz oddíl 4), mezi jednotlivé slovesné lexémy. Z grafů na obrázku 3 je vidět nepřímo úměrný vztah mezi počtem lexémů a složitostí jejich valenčního chování – největší počet sloves má jeden valenční rámec (a jde tedy o monosémické lexémy), mnoho sloves má několik málo valenčních rámců; naopak malé množství sloves vykazuje velmi složité valenční chování (počty valenčních rámců pro jednotlivá slovesa vycházejí ze slovníku *VALLEX 2.5*).

Z těchto pozorování vyplývá, že náročná manuální příprava slovníku má své opodstatnění. Nejčastější slovesa vykazují nejsložitější valenční chování a jejich syntaktické charakteristiky se nedaří získat automatickým zpracováním z existujících dat. Je proto výhodné zpracovat je ručně a získat tak jejich spolehlivý a konzistentní popis. Tím zároveň rychle roste pokrytí výskytu zpracovaných sloves v textech. Se snižující se frekvencí slovesa v korpusu postupně klesá jeho (průměrná) složitost, pro velmi řídká slovesa lze tedy předpokládat jednoduché valenční chování, které bude možné získat pomocí (polo)automatických metod. Ovšem otázka, kde je hranice takto jednoduchých sloves, zůstává zatím otevřená. Z grafu na obrázku 2 je například vidět, že slovesa z poslední skupiny, tedy okolo tisící pozice vzhledem k frekvenci v ČNK, mají průměrně přibližně 1,5 valenčního rámce. Přesto se i zde objevují slovesa

[7] Přestože jsou ve slovníky *VALLEX, verze 1* vidové protějšky zachyceny zvlášť, je vhodné zpracovávat je společně v rámci jediného lexému, viz zde zejména oddíl 4. Při výběru sloves ke zpracování byly proto k frekventovaným slovesům přidány též jejich vidové protějšky. Tato méně četná slovesa vykazují podobnou složitost jako jejich četnější vidové protějšky, i když mají nižší frekvenci v ČNK.

[8] Počet valenčních rámců pro daný lexém je dán součtem valenčních rámců pro jednotlivé vidy.



Obrázek 3: Počet lexémů ve slovníku *VALLEX, verze 2* vzhledem k počtu lexikálních jednotek (vlevo počet lexémů v lineárním měřítku; vpravo počet lexémů v logaritmickém měřítku). Grafy jsou založeny na počtech LU ve slovníku *VALLEX 2.5* – lexém opět sdružuje vidové protějšky; je-li LU sdílena dvěma, příp. více vidovými protějšky, je započítána pouze jednou, což odpovídá struktuře této verze slovníku.

velmi složitá – např. sloveso *vytáhnout* ^{pf} (pozice 1 000 v ČNK, řazeno podle klesající frekvencce) má 13 valenčních rámců pro nereflexivní lemma a další 3 rámce pro lemma reflexivní, z toho 9 pro idiomatická užití; 10 rámců z celkového počtu sdílí se svým nedokonavým protějškem *vytahovat* ^{impf} (pozice 1 803 v ČNK); vysoký počet valenčních rámců, které reprezentují jednotlivé lexikální jednotky, je v tomto případě způsoben homonymní předponou *vy-* a dále řadou rámců pro idiomy (počty vycházejí ze slovníku *VALLEX 2.5*).

První zveřejněná verze slovníku: *VALLEX, verze 1*

Intenzivní práce na ručním budování slovníku *VALLEX* začaly v roce 2001. Byla vypracována podrobná zpráva [51], shrnující různé přístupy k popisu valence ve světě (angličtina, němčina, polština, slovenština, ruština, bulharština a japonština) i u nás (především teorie větných vzorců, viz [9], teorie valence ve Funkčním generativním popisu, viz [58, 59, 60]). Tato zpráva popisuje jednotlivé ve slovníku zachycované jevy (zejména valenční rámce a jejich rozšíření o kvazivalenční a typická doplnění, ale i další jevy související s valencí, jako je reflexivita a reciprocita, kontrola či vidové dvojice). Dále jsou v ní představeny nástroje, které byly vyvinuty pro potřeby anotátorů, jmenujme především textový editor s kontrolou syntaxe (‚syntax highlighting‘), rozhraní pro vyhledávání v elektronických zdrojích (zejména ve slovnících *SSJČ* [19], *BRIEF* [56], *Slovesa pro praxi* [87] a ve vzorku ČNK) a v neposlední řadě též rozhraní pro pokročilé vyhledávání v datech slovníku *VALLEX*. Tyto nástroje jsou spolu s XML datovou strukturou slovníku podrobně popsány v dizertaci Z. Žabokrtského [90].

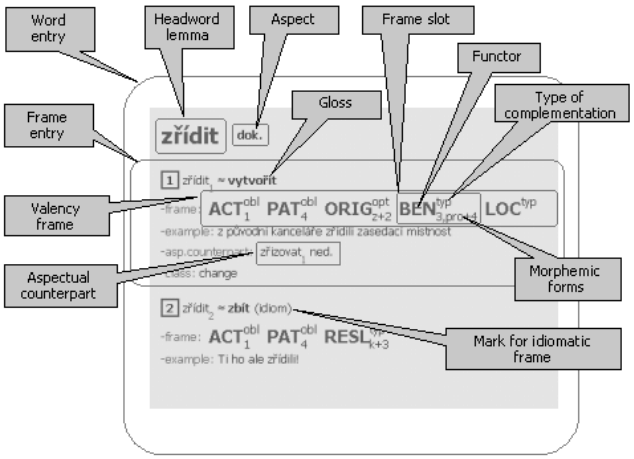
Prvním veřejně dostupným výsledkem projektu budování slovníku (a vůbec prvním elektronickým slovníkem s valenční reprezentací českých sloves) byl slovník *VALLEX, verze 1.0*, který byl zpřístupněn na webových stránkách Ústavu formální a aplikované lingvistiky MFF UK v roce 2003.[9]

VALLEX 1.0 popisoval valenční chování zhruba 1 400 českých sloves – k jádru

[9] `http://ufal.mff.cuni.cz/vallex/1.0/`

slovníku *Vallex-00* bylo přidáno 1 000 nejčastějších sloves a jejich vidových protějšků (podle frekvence v ČNK, s výjimkou slovesa *být*). Těmto slovesům odpovídá téměř 2 500 valenčních rámců. Slovesa obsažená ve slovníku *VALLEX 1.0* pokrývají přibližně 56,4% výskytů sloves v daném subkorpusu ČNK (na sloveso *být* a modální slovesa připadá dalších 28,5%), viz též graf na obrázku 1.

Struktura dat ve slovníku *VALLEX 1.0* je podrobně popsána v příspěvku [91] a v Nápovědě k HTML formátu slovníku. Na nejvyšší úrovni je slovník tvořen slovesnými hesly, která jsou reprezentována jednotlivými slovesnými lemmaty (morfologické varianty lemmatu jsou popsány ve společném hesle). Slovníkové heslo se skládá ze záznamů pro jednotlivé valenční rámce (‚frame entry', viz obrázek 4, zachycující typicky sloveso v jednom z jeho významů). Kromě vlastního valenčního rámce tyto záznamy obsahují povinné atributy (glosa, příklad) a nepovinné atributy (zejména kontrola, syntakticko-sémantická třída a odkaz na vidový protějšek).



Obrázek 4: Struktura hesla ve slovníku *VALLEX, verze 1* (převzato z Nápovědy pro HTML formát slovníku)

Při budování slovníku se od počátku kladl velký důraz na co nejširší možnost využití slovníku jak pro počítačové zpracování češtiny, tak pro člověka jako uživatele jazyka. Proto byl zveřejněn ve třech formátech:

HTML formát. *VALLEX 1.0* jako webová aplikace umožňuje snadné vyhledávání ve slovníku podle různých aspektů (vedle základního abecedního uspořádání sloves též např. podle funktorů[10], forem doplnění, syntakticko-sémantických tříd či kontroly).

PDF formát. *VALLEX 1.0* ve verzi pro tisk (byl vydán též jako technická zpráva [48]).

[10] Funktory označují typy syntakticko-sémantických vztahů mezi slovesem a jeho doplněním, tedy např. ACT pro aktor, PAT pro patient, DIR3 pro doplnění směru-kam; přehled funktorů lze nalézt např. v práci [47].

XML formát. XML je primárním formátem slovníku *VALLEX 1.0*, vhodným pro aplikace a pro pokročilé vyhledávání.

Zpracování velkého množství dat potvrdilo, že přijatá koncepce slovníku je nosná a metodologie zpracování slovesných hesel vyhovuje požadavkům na efektivní a konzistentní lexikografickou práci (testování konzistence a úplnosti slovníku *VALLEX 1.0* bylo popsáno v příspěvku [49]). To se také odrazilo na zájmu uživatelů o slovník *VALLEX 1.0*. Proto bylo možné přistoupit ke kvantitativnímu i kvalitativnímu rozšiřování slovníku, jehož prvním výsledkem byla pracovní verze slovníku *VALLEX 1.5*.

VALLEX 1.5 obsahoval 2 500 sloves (přibližně 6 000 valenčních rámců), včetně slovesa *být* a modálních sloves. Tato verze, která byla zpřístupněna zejména uživatelům z Ústavu formální a aplikované lingvistiky, sloužila pro rozsáhlé testování zvolených formátů. Velký důraz byl kladen zejména na konzistenci a systematičnost zachycení jednotlivých jevů popisovaných ve slovníku; představovala hodnotnou zpětnou vazbu, která odhalovala problematická rozhodnutí a nesystematická řešení, a tím umožňovala další zvýšení kvality slovníku. *VALLEX 1.5* je popsán v článku [40]. Zde jsou prezentovány i první výsledky experimentů s automatickým přiřazováním valenčních rámců jednotlivým výskytům sloves v textu (více viz závěrečné poznámky v oddíle 6).

3 Teoretické aspekty valence a jejich promítání do koncepce slovníku

Teoretický základ pro vytváření valenčního slovníku je dán Funkčním generativním popisem češtiny (FGD, Functional Generative Description), viz zejména [77, 81], který je charakteristický svým závislostním stratifikačním přístupem k popisu jazyka. V rámci tohoto přístupu byla teorie valence rozvíjena od sedmdesátých let; zásadní shrnutí lze nalézt zejména ve stati [60]. Valenční rámec se v tomto pojetí skládá z aktantů, tedy vnitřních doplnění (obligatorních i fakultativních), a z obligatorních volných (adverbiálních) doplnění.

Při budování slovníku *VALLEX* byla beze zbytku přejata koncepce FGD. Její koncepty jsou zde aplikovány na velké množství dat – zatímco v teoretických pracích byly vždy popisovány především základní významy sloves, nyní jsou zpracovávány celé slovesné lexémy. Tyto aspekty s sebou nesou nutnost zpřesňovat funkční kritéria stanovená v rámci teoretického výzkumu a formulovat kritéria další.

Základní teoretické aspekty budování valenčního slovníku jsou (kromě dalších témat) popsány v již výše zmíněném rozsáhlém článku [39]. Je zde nastíněn zejména koncept kvazivalenčních doplnění a dále je zde věnována pozornost vyčleňování jednotlivých významů sloves. Obě tato témata byla dále zkoumána na velkém množství korpusových dat – jejich podrobnější popis lze nalézt v článku [42].

Kvazivalenční doplnění

Již v dizertaci [37] a poté v knižní publikaci [38] byl zaveden pojem kvazivalence, kterým se dosavadní pojetí valence jako souboru doplnění vnitřních (aktantů) a obligatorních volných rozšiřuje i na doplnění často užívaná, ‚ustálená'. Na základě velkého

množství dat zpracovávaných jednak při pracích na slovníku, jednak též při anotaci PDT se ukázalo, že existuje skupina doplnění, která sice nesplňují striktní kritéria pro vnitřní doplnění, přitom jsou však tato doplnění lexikálně vázaná. Svými vlastnostmi se tedy blíží aktantům:

- jejich morfématická podoba je určována požadavky řídícího slovesa;
- rozvíjejí omezenou třídu sloves;
- jako doplnění jednoho slovesa se nemohou opakovat.

Svými dalšími charakteristikami však odpovídají volným doplněním (viz též [42], kde jsou kritéria rozlišující aktanty a volná dolnění formulována):

- jsou sémanticky homogenní;
- jsou převážně fakultativní;
- nepodléhají posouvání.

Jako kvazivalenční doplnění bylo klasifikováno doplnění záměru pro určitá slovesa pohybu (INTT, např. *Petr mamince doběhl nakoupit*), doplnění překážky pro podtřídu sloves kontaktu (OBST, např. *Chlapec zakopl o kořen*) a doplnění rozdílu pro slovesa vyjadřující změnu (DIFF, např. *Hodnota akcií stoupla o 100 %*). Dále bylo jako kvazivalenční doplnění stanoveno doplnění mediátoru (MDT, např. *Když jsem odcházel, zatahal mě soused za rukáv*), toto doplnění však zatím nebylo uplatněno ve slovníkových datech.

Otevřenou otázkou prozatím zůstává klasifikace adresátu (ADDR) a původu (ORIG), které jsou v rámci ‚klasické' teorie FGD považovány za doplnění vnitřní, svou sémantickou homogenností se však blíží doplněním kvazivalenčním.

Nově navržený koncept kvazivalenčních doplnění vhodným způsobem obohacuje tradiční dělení slovesných doplnění, umožňuje jejich jemnější klasifikaci a přiměřené zachycení doplnění na hranici vnitřních a volných doplnění.

Specifikace významu a rozlišení jednotlivých významů sloves

V koncepci slovníku *VALLEX* valenční rámce (tedy soubor valenčních pozic, které jsou charakterizovány funktorem, zachycujícím syntakticko-sémantický vztah doplnění ke slovesu, možnými morfématickými formami a obligatorností) zhruba odpovídají jednotlivým významům daného slovesa (problematika tzv. alternací je zmíněna níže).[11]

Pro vyčleňování jednotlivých významů daného slovesa neexistují všeobecně přijatá testovatelná kritéria, přechod od jednoho významu k druhému je v řadě případů pozvolný. Ve slovníku *VALLEX* je při jejich rozlišování kladen důraz na syntaktická kritéria, zejména na podobu valenčního rámce. Přitom se ovšem přihlíží též k sémantice. V práci [42] jsou formulovány dva základní principy pro vyčleňování jednotlivých valenčních rámců:

[11] Valenční rámce jsou zde chápány jako základní syntakticko-sémantická informace charakterizující jednotlivé lexikální jednotky.

- Každá změna ve valenčním rámci (s výjimkou možných variant morfématického vyjádření jednotlivých funktorů) vede k vyčlenění nového valenčního rámce (srov. též s prací věnovanou alternačnímu modelu slovníku [50]).
- Každá signifikantní změna významu s sebou nese nutnost vyčlenění nového valenčního rámce slovesa.

Problematika vztahu valenčních rámců a jednotlivých významů sloves je podrobněji rozvedena v práci [47] pojednávající o současné struktuře slovníku, viz níže.

Zpracovávání sloves s podobnými sémantickými vlastnostmi

K tématům, které jsou již řadu let široce diskutována, patří vztah syntaktických a sémantických vlastností sloves, viz zejména knihy [35] a [36]. Při zpracovávání slovníkových hesel se osvědčilo zkoumání valence celých skupin sloves, která mají obdobné sémantické vlastnosti. Protože se ukázalo, že nelze jednoduchým způsobem adaptovat žádnou z existujících klasifikací sloves,[12] byl vytvořen návrh zhruba dvaceti velmi hrubých syntakticko-sémantických skupin, které spojují slovesa s podobným, příp. zcela stejným syntaktickým chováním. Tyto skupiny jsou velmi dobrým východiskem pro další podrobné zkoumání syntaktických i sémantických vlastností sloves – uveďme např. rozbor některých sloves výměny v článku [41] či rozbor vybraných prefigovaných sloves pohybu v příspěvku [43]. Další velmi zajímavou skupinou sloves, která je zkoumána v návaznosti na *VALLEX*, je rozsáhlá skupina sloves mluvení – např. v článku [29] se navrhuje vyčlenění dalších podtypů této třídy na základě větné modality závislé klauze (povrchově vyjádřené různými spojkami uvozujícími tuto klauzi) na slovesa povahy oznamovací, imperativní a tázací; stať [30] řeší rozpad tzv. tématu a dikta.

Alternační model slovníku

Důraz na syntaktická kritéria při zkoumání valence vede k vyčleňování zvláštních valenčních rámců pro taková užití sloves, která se liší syntaktickou strukturací, i když mají stejný nebo velmi blízký význam; např. dvojice *naložit vůz*$_{PAT}$ *senem*$_{EFF}$ vs. *naložit seno*$_{PAT}$ *na vůz*$_{DIR3}$ či *vyběhnout kopec*$_{PAT}$ vs. *vyběhnout na kopec*$_{DIR3}$ mají různé valenční rámce, a tudíž jim budou odpovídat různé lexikální jednotky, přesto by bylo vhodné zachytit jejich významovou blízkost.

Podívejme se v této souvislosti na subjektové diateze sloves. Různé diateze jsou též charakteristické různou strukturací věty, přitom není obvyklé vyčleňovat pro jednotlivé diateze zvláštní valenční rámce – vzhledem k velké pravidelnosti v morfématickém vyjádření diatezí by bylo explicitní vyčleňování samostatných valenčních rámců redundantní (stačí tedy informace o možnosti slovesa realizovat danou diatezi). A to přesto, že různé teoretické koncepce se mohou lišit v názoru na to, zda dvě věty lišící se pouze v diatezi mají stejný význam (např. věty se stejným lexikálním obsazením, ale s aktivní a pasivní slovesnou formou).

[12] Jmenujme zde alespoň již zmíněný přístup B. Levinové [35], který byl využit např. v projektu VerbNet, `http://wordnet.princeton.edu/`, či sémantické třídění sloves v projektu FrameNet, `http://framenet.icsi.berkeley.edu`. Pro češtinu je inspirativní zejména třídení sloves v monografii [9].

Obdobným způsobem lze přistoupit i k popisu syntakticky různě strukturovaných užití slovesa, pokud jsou tyto strukturní (a případně i významové) změny natolik systémové, že se dají zachytit pomocí pravidel. Lexikální jednotky popisující takto pravidly provázaná slovesné významy označujeme jako slovesné *alternace*.[13]

Alternační model slovníku, jehož logická struktura byla navržena v dizertaci Z. Žabokrtského [90], umožňuje systematické a ekonomické zachycování pravidelných změn významu jednotlivých sloves.

Alternační model slovníku je charakterizován dvěma komponentami, vlastní slovníkovou komponentou a komponentou gramatickou. *Slovníková komponenta* se skládá z lexémů, které sdružují jednotlivé lexikální jednotky, viz zejména [50]. *Gramatická komponenta* potom popisuje pravidelné změny pro jednotlivé alternace, a to na třech úrovních:

- případná změna slovesné formy;
- případné změny ve valenčním rámci (tedy změny v počtu valenčních doplnění a v jejich funktorech, změny v obligatornosti doplnění a změny v jejich morfématické realizaci);
- případné změny významu.

V příspěvku [50] jsou představeny i základní typy alternací – tzv. *syntaktické alternace* (zejména různé typy diatezí a reciprokalizace, které jsou bohatě zpracovány v českých mluvnicích a teoretických statích, viz např. [10, 13, 61]) a *sémantické alternace* (např. rušení/vytváření souvýskytu, viz též [8]) a jejich vzájemné vztahy.

Při provázání jednotlivých lexikálních jednotek alternačními pravidly lze valenční informaci ve slovníku zobrazovat v různě kompaktních verzích, například v závislosti na typu aplikace, pro kterou je daný slovník určen. Je-li slovník určen pro člověka, lingvistu či uživatele jazyka, je vhodné určité typy vztahů vyjadřovat implicitně (např. informace o možných diatezích v rámci jediné lexikální jednotky), jiné vztahy naopak explicitně (zdá se např. vhodnější zachycovat sémantické diateze jako dvě lexikální jednotky spojené odkazem).

Poznamenejme, že tento alternační model slovníku byl rozpracován na úrovni formální koncepce slovníku [50] a byla připravena jeho implementace [90]. Do dat slovníku *VALLEX* zatím alternační model promítnut nebyl, viz též závěrečné poznámky v oddíle 6.

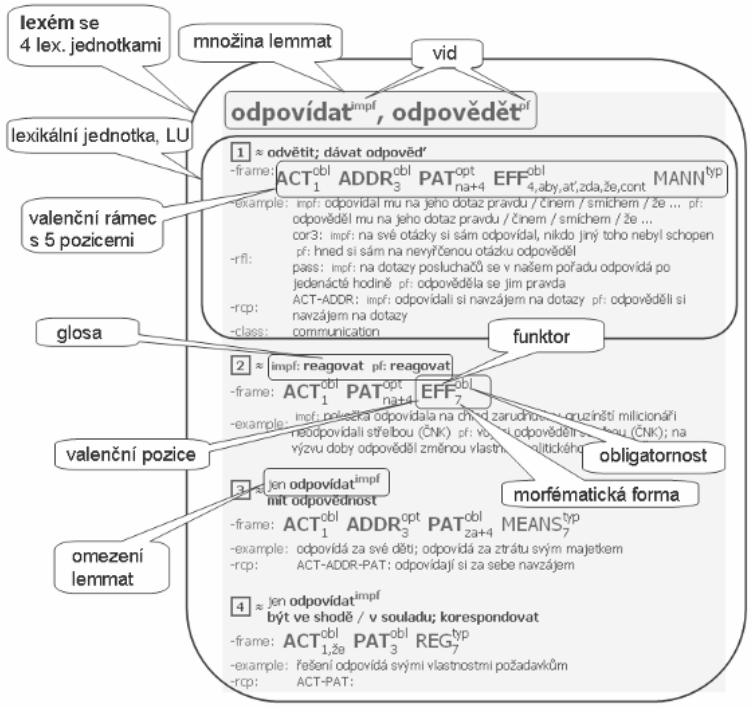
4 Současná koncepce valenčního slovníku: *VALLEX, verze 2*

V první verzi slovníku, označené zde jako *VALLEX 1.0* (a též v jeho kvantitativním rozšíření *VALLEX 1.5*), jsou vidové protějšky sloves zachyceny jako samostatné slovníkové položky (propojené pouze odkazem). Takovéto pojetí vidových protějšků

[13] Přínosnou inspirací pro zkoumání různých typů alternací je práce B. Levinové [35], která též termín ‚alternace' zavedla. Na základě bohatého souboru alternací nejrůznějších typů pak navrhla systém sémantických tříd. Přestože se její práce soustřeďuje na anglická slovesa, lze najít řadu paralel v syntaktickém chování českých sloves.

jako samostatných jednotek není v souladu s teoretickými východisky danými FGD, kde je vid považován za gramatickou kategorii a vidové protějšky (v běžné terminologii vidové dvojice) za různé realizace jednoho lexému, viz [64].

Tomuto pojetí vidu lépe odpovídá současná verze slovníku. *VALLEX, verze 2*[14] zachycuje valenční chování vidových protějšků v rámci jediného lexému, kterému odpovídá vždy jedno slovníkové heslo. Struktura slovníku *VALLEX, verze 2* je podrobně popsána v úvodu k tištěné verzi slovníku [47] a též jako Nápověda k HTML formátu na webových stránkách slovníku. Na nejvyšší úrovni je slovník *VALLEX* tvořen lexémy – lexémem přitom rozumíme abstraktní jednotku, která v sobě spojuje formální složku (množinu všech slovních forem reprezentovanou možnými lemmaty) i významovou složku, reprezentovanou jednotlivými lexikálními jednotkami (LU, lexical unit), česky označovanými též jako lexie nebo základní lexikální jednotky. Struktura slovníkového hesla je přiblížena na obrázku 5.



Obrázek 5: Struktura hesla ve slovníku *VALLEX, verze 2* (převzato z Nápovědy pro HTML formát slovníku)

VALLEX, verze 2 popisuje chování 2 730 českých lexémů, které zahrnují 6 460 lexikálních jednotek – ‚daných sloves v daném významu'. Pokud bychom počítali dokonavá a nedokonavá slovesa zvlášť, dostali bychom se k počtu 4 250 sloves. Hlavním kritériem výběru sloves ve slovníku *VALLEX* byla jejich frekvence v ČNK – v prvním

[14] Nadále zde nebudeme důsledně rozlišovat verze slovníku *VALLEX 2.0* a *VALLEX 2.5* – obě tyto verze mají stejnou strukturu a popisují stejný počet sloves, verze *2.5* byla podrobena rozsáhlým poloautomatickým kontrolám správnosti a konzistence při přípravě tištěné verze slovníku.

kroku bylo vybráno přibližně 2 500 nejčastějších slovesných lemmat, posléze byl tento výběr doplněn tak, aby slovník obsahoval ke každému slovesu i jeho vidové protějšky.[15]

Slovník poskytuje informace o valenční struktuře českých sloves v jejich jednotlivých významech, které charakterizuje pomocí glos a příkladů. Pro jednotlivá valenční doplnění uvádí *VALLEX* možná morfématická vyjádření, pokud jsou jejich formy dány slovesnou rekcí. Kromě těchto základních údajů uvádí i některé další syntaktické, případně syntakticko-sémantické charakteristiky, jako je vlastnost kontroly, typ reflexivního užití, možnost recipročního užití či syntakticko-sémantická třída slovesa.

Formální datová struktura slovníku, jeho technologické a technické aspekty jsou popsány v dizertaci Z. Žabokrtského [90].

VALLEX 2.5 byl zveřejněn na webových stránkách Ústavu formální a aplikované lingvistiky Matematicko fyzikální fakulty Univerzity Karlovy v Praze koncem roku 2007.[16] Od jara 2008 je k dispozici též jeho tištěná verze, kterou vydalo Karolinum, nakladatelství Univerzity Karlovy v Praze [47].

5 Formální modelování přirozeného jazyka: valence jako základní syntaktická informace

Valenční charakteristika slovesa představuje základní syntaktickou informaci, která určuje strukturu věty, v rámci FGD viz např. [78, 79]; někdy se proto též mluví o *valenční (závislostní) syntaxi*. Proto se valence jako základní závislostní vztah zásadním způsobem uplatňuje i při formálním modelování syntaktické struktury přirozeného jazyka a jeho syntaktické analýzy.

Redukční analýza pro valenční (závislostní) syntax

Jako výchozí metodu při popisu syntaktické struktury přirozených jazyků, zejména jazyků s volným slovosledem, využíváme metodu *redukční analýzy*.[17] Tato elementární metoda je podrobně popsána v článku [44]. Redukční analýza spočívající v postupném zjednodušování analyzované věty umožňuje formálně stanovit závislostní vztahy mezi jednotlivými členy věty (ať už jednotlivými slovy či slovními spojeními). Přitom je důležité, že při redukční analýze – na rozdíl od složkových přístupů, kde je základní operací rozklad věty na souvislé úseky odpovídající jednodušším strukturám-frázím – lze při určování závislostních vztahů mezi slovními spojeními do jisté míry abstrahovat od jejich slovosledných pozic ve větě, aniž by však slovosledná informace byla opomíjena. Princip redukční analýzy můžeme shrnout do následujících pozorování:

[15] Takto jsou zpracovány vidové protějšky tvořené sufixálně a řídké supletivní páry – praktickým důvodem je nejednoznačnost při určování prefigovaných vidových protějšků; k možnosti propojení prefigovaných dokonavých sloves s jejich neprefigovanými nedokonavými protějšky viz též závěrečné poznámky v oddíle 6.

[16] `http://ufal.mff.cuni.cz/vallex/2.5/`

[17] Zdůrazněme zde, že nám jde o *formální* model analýzy, nikoli o model *psycholingvistický*, který by měl přiblížit proceduru porozumění vět přirozeného jazyka v našem vědomí.

1. Skutečnost, že nějaké slovo či slovní spojení lze z věty vypustit (a přitom získat větu jednodušší) znamená, že toto slovo či slovní spojení závisí na některém slově (slovním spojení) ze zkrácené věty.
2. Dvě slova či slovní spojení lze postupně vypustit v libovolném pořadí, právě když jsou vzájemně nezávislá.
3. Některá slovní spojení je (vzhledem k principům uvedeným níže) nutno vypouštět v jednom kroku – i v tomto případě je při závislostní analýze obvyklé určovat závislosti. Při určování řídících členů takových spojení je potřebné dále zkoumat pravidla pro jednotlivé jazykové jevy.

Při postupné redukci analyzované věty je nutné uplatňovat některé základní principy:

- zachování syntaktické správnosti věty;
- zachování lemmatu (slovníkového hesla) a vybrané morfologické značky (souboru morfologických kategorií, které charakterizují daný výskyt slova);
- zachování významu původních slov ve větě (zde je význam reprezentován valenčním rámcem);
- zachování významové úplnosti věty (dále zejména zachování informace o všech aktantech a obligatorních valenčních doplnění všech neredukovaných slov ve větě).

Metoda redukční analýzy tedy umožňuje odvodit závislostní, zejména valenční vztahy ve větě z možných pořadí redukcí jednotlivých slov a slovních spojení. To je podstatné zejména pro jazyky jako je čeština, kde závislostní strukturu nelze odvodit primárně ze slovosledu. Slovosled zde odráží aktuální členění věty, tedy nese významovou informaci [17]. Změna slovosledu s sebou nemusí nutně nést změnu závislostní struktury věty, přitom však nelze věty lišící se pouze slovosledem považovat za synonymní. Redukční analýza tedy dovoluje zkoumat závislostní vztahy a slovosled do určité míry nezávisle.

Stať [44] se soustřeďuje na vyjasnění vztahů mezi redukční analýzou a reprezentací využívající formalismus závislostní gramatiky, viz např. [67]. Ukazuje, že pomocí kroků 1. a 2. lze zpracovávat endocentrické konstrukce, zejména (autosémantická) slova a jejich fakultativní volná doplnění – slovo či slovní spojení, které se při redukci chová jako řídící, odpovídá rozvíjenému slovu či spojení ve větě, zatímco slovo či slovní spojení v redukci závislé odpovídá slovu či slovnímu spojení rozvíjejícímu.

Dále tento příspěvek analyzuje tzv. *redukční komponenty*, slovní spojení tvořená slovy, která je nutno vypouštět v jediném redukčním kroku, viz bod 3. Redukční komponenty odpovídají exocentrickým konstrukcím – modelují například formémy, tedy slovní spojení tvořící jednotlivé větné členy (např. předložkové skupiny či analytické slovesné tvary, viz [81, 78]); pro výběr řídícího členu se v těchto případech stanovují pravidla spíše technického charakteru, která se v různých aplikacích mohou lišit. Daleko zajímavějším jevem, který lze modelovat pomocí redukčních komponent, je ovšem valenční struktura sloves a dalších autosémantických slov. Při určování směru závislosti mezi slovy s valenčními požadavky a jejich doplněními se potom uplatňuje princip analogie na úrovni slovních druhů navržený v monografii [81].

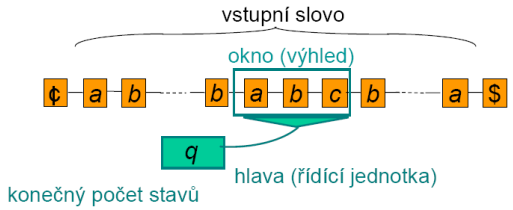
Restartovacích automaty jako model redukční analýzy

Závislostně orientovaný Funkční generativní popis byl původně formálně modelován jako generativní systém pomocí soustavy sériově řazených zásobníkových a konečných automatů-převodníků, viz zejména [77, 82, 72]. Model vlastní generativní složky FGD, která vytváří významovou (tektogramatickou) reprezentaci věty, byl později podrobně popsán v článku [65]. Také tento model byl realizován jako zásobníkový automat.

V osmdesátých letech byl potom navržen systém překladových schémat, který umožňoval interpretaci v obou směrech – jako generativní i analytický systém [66]. Podobně je FGD představen i v učebnici [80], kde je modelován jako kompozice několika bezkontextových jazyků odpovídajících jednotlivým rovinám popisu.

Redukční analýza se stala významnou motivací pro nový formální model FGD založený na konceptu *restartovacích automatů*. Zde podáme jen základní neformální přiblížení restartovacích automatů, jejich formální popis a podrobnou typologii lze najít v bohaté bibliografii věnované těmto automatům, jejich vlastnostem a hierarchiím jimi přijímaných jazyků, např. [25, 55] a tam citované publikace. Redukční analýzu dobře modelují restartovací automaty, které pracují s tzv. vlastním jazykem (jazyk vstupní věty) a s charakteristickým jazykem (vstupní jazyk obohacený o gramatické kategorie popisující strukturu věty), viz zejména [53, 54, 70].

Konkrétní model restartovacího automatu je formálně definován v článku [45]. Restarovací automat modelující redukční analýzu (dále automat typu RA) je formální zařízení – nedeterministický matematický stroj – s konečně stavovou řídící jednotkou a s hlavou s omezeným výhledem, která čte a zpracovává větu na pružné pásce (ohraničené speciálními symboly) nad konečným slovníkem, viz obrázek 6.



Obrázek 6: Schéma restartovacího automatu typu RA

Tento automat začíná výpočet nad vstupní větou v daném počátečním stavu a s hlavou umístěnou na levém okraji pásky. Během výpočtu provádí – podle své přechodové funkce – následující operace:

MVR/MVL: posuny doprava / doleva;

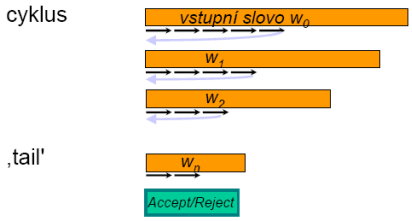
Rewrite(v): přepisovací kroky, které zkracují slovo na pracovní pásce (tj. přepisují slovo ve výhledu kratším slovem v);

Restart: hlava automatu se posune na levý okraj pásky a přepne se do počátečního stavu;

Accept/Reject: automat přijme / zamítne slovo na pásce.

Restarovací automat typu RA pracuje v cyklech (obrázek 7) – zpracovává vstupní větu (podle přechodové funkce automatu hlava čte slova ve výhledu, posouvá se po

pásce doprava i doleva a přepisuje / zkracuje větu na pásce), dokud větu na pásce nepřijme či neodmítne, nebo dokud nedojde k operaci ‚Restart'. Při této operaci hlava automatu ‚zapomene' svou pozici na pásce i svůj vnitřní stav a začíná zpracovávat (již částečně zpracovanou) větu od začátku v novém cyklu. (Přitom požadujeme, aby před prvním restartem i mezi každými dvěma restarty byla věta přepsána / zkrácena.)



Obrázek 7: Výpočet restartovacího automatu v cyklech

Základním předpokladem pro modelování redukční analýzy je vlastnost *zachování chyby* – při přijímajícím výpočtu restartovacího automaty typu RA je před i po každém jeho cyklu na pásce věta z charakteristického jazyka (tedy z vlastního jazyka slovních forem obohaceného o gramatické kategorie popisující stavbu věty). Pro vlastní modelování redukční analýzy se obvykle požaduje ještě silnější vlastnost, tzv. *zachování správnosti* – tato vlastnost zaručuje, že každý výpočet automatu typu RA nad větou z charakteristického jazyka RA je přijímající.

Modelování redukční analýzy (a na jejím základě též syntaktické analýzy) pomocí restartovacích automatů odpovídá v řadě ohledů lépe paradigmatu FGD než starší modely založené na zásobníkových automatech:

- Restartovací automaty typu RA adekvátním způsobem modelují syntaktické vztahy dané valenčními charakteristikami autosémantických jednotek. Umožňují totiž několik přepisování v jednom cyklu – touto technikou lze redukovat celé redukčních komponenty odpovídající vždy slovu a jeho valenčním doplněním. Zpracování jednoho slovesa, substantiva, adjektiva nebo adverbia a jeho valenčních doplnění je tedy modelováno jedním cyklem výpočtu. Tímto se podstatným způsobem liší od modelů založených na zásobníkových automatech, které modelovaly *syntaktické dvojice* řídící – závislé slovo, nikoli tedy plnou valenční strukturu jako jádro věty, jak ji chápe valenční syntax.
- Restartovací automaty typu RA dovolují přirozeně zachytit myšlenku *lexikalizace*, tedy přístupu vlastního závislostnímu popisu jazyka, který shromažďuje základní jazykové informace ve slovníku (odkažme zde alespoň na kategoriální gramatiky, viz [2], a na koncepci ‚Lexicalized Tree Adjoining Grammar', viz např. [1], navazující na ‚Tree Adjoining Grammar' A. Joshiho [26]; v rámci FGD potom viz např. [78]).
- Restartovací automaty typu RA zároveň dobře zachycují nelokální chování jazyků s volným slovosledem – přepisovací kroky nejsou v těchto obecných modelech automatů omezeny na okolí jedné slovosledné pozice [69, 70], mohou proto v jednom cyklu zpracovávat i slova v povrchové realizaci věty vzdálená.

- Postup restartovacího automatu v cyklech vhodným způsobem modeluje rekurzivní vlastnosti jazyka – nejprve jsou zpracovávány nejhlouběji vnořené jazykové konstrukce a tím je analyzovaná věta zjednodušena; následuje zpracování jazykových konstrukcí vnořených v takto zjednodušené větě; po každém zjednodušení začíná nový cyklus. Výpočet pokračuje, dokud není získána základní predikační struktura věty, která je už přijímaná bez restartu (v ‚tailu' výpočtu, viz obrázek 7), nebo dokud není (zjednodušená) věta zamítnuta.

Funkční generativní popis jako formální překlad

Po formálním systému pro popis přirozeného jazyka požadujeme (viz [66, 80]), aby popisoval množinu správných vět jazyka, množinu možných významových (tektogramatických) reprezentací vět v daném jazyce a vztahy mezi těmito dvěma množinami, které odpovídají vztahům reprezentace (a zachycují tedy synonymii a homonymii v jazyce).

V již zmíněném příspěvku [45] je definován 4-úrovňový redukční systém, který představuje nový formální rámec pro modelování FGD založený na principech redukční analýzy.

Restartovací automat modelující FGD, dále M_{FGD}, zpracovává věty nad charakteristickým slovníkem Σ, který sestává jednak ze všech slovních forem popisovaného přirozeného jazyka, jednak z gramatických kategorií popisujících strukturu věty. Stratifikační přístup FGD se zde promítá do rozdělení tohoto charakteristického slovníku Σ na (pod)slovníky pro jednotlivé jazykové roviny:

Σ_0 ... slovník sestávající ze všech správných slovních forem jazyka; vlastní jazyk automatu M_{FGD};

Σ_1 ... slovník pro morfologickou analýzu (popisuje lemmata a jejich morfologické značky);

Σ_2 ... slovník reprezentující povrchovou syntax;[18]

Σ_3 ... slovník obsahující jednotky popisující významovou charakteristiku autosémantických slov (zejména lexikální a valenční informaci, funktory, gramatémy a aktuální členění).

V článku [45] je též specifikován charakteristický jazyk restartovacího automatu M_{FGD}, který zachycuje modelovanou větu i její analýzu na jednotlivých rovinách (pomocí slovníků $\Sigma_0, \Sigma_1, \Sigma_2$ a Σ_3).[19] Dále je zde podrobně ilustrován výpočet automatu M_{FGD} při zpracování konkrétních českých vět. Pozornost je věnována dvěma základním lingvistickým jevům. Je to především zpracování valenčních a volných doplnění tak, aby byl zachován princip zachování úplnosti vstupní věty, tedy jeden ze základních principů redukční analýzy dovolující adekvátně postihnout vlastnosti tektogramatické reprezentace věty. Dalším zevrubně zpracovaným jevem je zachycení povrchového i hloubkového slovosledu (včetně zachycení neprojektivních konstrukcí, které se mohou objevit v povrchové reprezentaci věty, viz zejména [23, 92, 18]).

[18] Otázku teoretické adekvátnosti této roviny FGD zde ponecháváme stranou.

[19] Slovník Σ_0 je tedy slovník vlastního jazyka automatu M_{FGD}, charakteristický slovník $\Sigma = \Sigma_0 \cup \Sigma_1 \cup \Sigma_2 \cup \Sigma_3$ je slovníkem charakteristického jazyka automatu M_{FGD}.

Restartovací automat M_{FGD} přijímá právě všechny správně vytvořené věty modelovaného přirozeného jazyka spolu s jejich (zjednoznačněnou) reprezentací na všech rovinách; zamítá věty, které nepatří do tohoto jazyka nebo nemají správnou reprezentaci na některé z rovin popisu.

Formální vztah mezi projekcí zpracovávané věty do roviny reprezentované slovníkem Σ_0 (tedy do vlastního jazyka M_{FGD}) a projekcí do roviny dané slovníkem Σ_3 definuje *charakteristickou relaci*. Tato relace modeluje vztahy reprezentace, tedy vztahy mezi množinou vět přirozeného jazyka a množinou významových (tektogramatických) reprezentací vět. Charakteristickou relaci můžeme též interpretovat jako překlad z jazyka správně utvořených vět do jazyka tektogramatické reprezentace (analýza) či jako překlad z jazyka tektogramatického do jazyka správných vět (syntéza).

Formální model FGD je dále rozpracován v přednášce [68], kde je tento systém začleněn mezi formální překladové systémy. Důraz je přitom kladen na propojení formálních modelů s jejich jazykovou (lingvistickou) náplní.

Rámec redukční analýzy (a jeho modelování restartovacími automaty) umožňuje formulovat podrobná pravidla pro zachycování jednotlivých jazykových jevů. Při zpracovávání valenčních informací a základních slovosledných konstrukcí (včetně povrchové neprojektivity) se lze omezit na restartovací automaty, kde se operace přepisování redukuje na vypouštění (tj. část pásky ve výhledu se přepisuje tak, že se některé symboly vypustí). Jiné konstrukce, například číslovkové konstrukce a zejména koordinační konstrukce, vyžadují obecnější model restartovacího automatu s přepisováním.

Poznamenejme, že koordinační a apoziční konstrukce byly zatím při formálním popisu založeném na restartovacích automatech ponechány stranou, neboť pracují s jednotkami, které mají složkový charakter [71, 78]. Významně tak překračují přímočaré pojetí závislostního přístupu. Nicméně v současné době jsou i tyto konstrukce postupně zpracovávány v rámci redukční analýzy a paradigmatu daného restartovacími automaty.

6 Závěrečné poznámky

Využití slovníku *VALLEX* pro budování dalších slovníků a pro aplikace v NLP

Při navrhování koncepce slovníku *VALLEX* byl kladen důraz na přesnost a lingvistickou adekvátnost popisu valence u velkého množství sloves. Slovníková hesla byla zpracovávána manuálně s přihlédnutím ke korpusovému i slovníkovému materiálu s následnou rozsáhlou automatickou, poloautomatickou i ruční kontrolou. Od samého počátku se předpokládalo využití slovníku *VALLEX* jak pro člověka jako uživatele jazyka, tak pro počítačové zpracování češtiny a pro další aplikační účely, jako např. strojový překlad, vyhledávání v textech apod.

- Valenční slovník *VALLEX* využívá přes 150 zaregistrovaných uživatelů, především z českých, ale i zahraničních univerzit. V rámci ČR bylo nejvíce licencí vyžádáno z pracovišť Matematicko-fyzikální a Filozofické fakulty UK,

Fakulty informatiky MU a Ústavu pro jazyk český AV ČR, z desítek licencí pro zahraniční instituce pak lze uvést např. Statní univerzitu v Ohiu (The Ohio State University), Univerzitu v Sársku (Universität des Saarlandes), Univerzitu v Záhřebu (Sveučilištu u Zagrebu) či francouzská pracoviště INALCO (Institut National des Langues et Civilisations Orientales) a LaLIC (Language, Logiques, Informatique, Cognition) na Univerzitě Paříž-Sorbona (l'Université Paris-Sorbonne).

Jmenujme zde dále alespoň některé aplikace, které využívají slovníková data nebo způsob jeho technologického zpracování.

- Datová struktura slovníku *VALLEX 1.0* a jeho technologické řešení (zejména formát dat, XML reprezentace, konverzní a validační skripty Z. Žabokrtského) jsou využívány pro budování slovníku *VerbaLex*, viz [21, 22, 20].
- Struktura slovníku *VALLEX* a zkušenosti získané při zpracování sloves byly využity při vytváření valenčního slovníku pro anglická slovesa *EngVallex*, viz [4], budovaného na základě slovníku *PropBank Lexicon* [32]. Zatímco slovník *PropBank Lexicon* byl využit při anotacích argumentové struktury sloves v korpusu PropBank [57], slovník *EngVallex* se využívá při manuální anotaci tektogramatické reprezentace angličtiny.
- Některé principy a zkušenosti získané při budování slovníku *VALLEX* byly využity též při přípravě Švédsko-českého valenčního slovníku, viz [5, 6].
- Valenční teorie byla primárně rozvíjena pro slovesa. Od valenčních vlastností sloves lze potom odvozovat závěry o valenčním chování deverbativních substantiv (a adjektiv). Deverbativní substantiva do určité míry dědí valenční rámec zdrojových sloves, viz zejména [62, 63]. Podstatný je přitom typ derivace – zda jde o derivaci syntaktickou nebo lexikální.
  Pilotní studie zabývající se možnostmi predikce substantivního rámce na základě typu derivace (zejména typ sufixu) a valenčního rámce fundujícího slovesa stanoveného ve slovníku *VALLEX 1.0* byla shrnuta v příspěvku [46], dále potom byla důkladně zpracována ve statích a v dizertaci V. Kolářové-Řezníčkové, viz zejména [33, 34].
- Náhodně vybraný vzorek 109 sloves ze slovníku *VALLEX 1.0* byl využit pro vytvoření korpusu *VALEVAL*, viz [3]. Pro tato slovesa bylo z ČNK vybráno vždy 100 vět, ve kterých jim byly ručně přiřazeny valenční rámce. Takzvaná zlatá data, *golden VALEVAL*, představují množinu vět, ve kterých se anotátoři shodli na přiřazeném významu. Dosažená mezianotátorská shoda okolo 75% (měřená pro dvojice anotátorů) odpovídá testům pro významné slovníky anglických sloves, např. *PropBank Lexicon* (ústní sdělení M. Palmerové).
  Korpus vět s jednoznačným rozlišením významů slov je základním předpokladem pro vývoj nástroje pro automatické přiřazování významu jednotlivým výskytům slov, tedy pro úkol známý jako ‚word sense disambiguation'.
- Korpus VALEVAL se stal základem pro trénování nástrojů pro automatické určování významu sloves, které využívají různých metod strojového učení, viz zejména [40, 76, 75] Ukázalo se, že zpracování sloves ve slovníku *VALLEX* je natolik konzistentní, že umožňuje natrénovat nástroj, který s úspěšností 77,2%

určí správný valenční rámec pro daný výskyt slovesa (oproti ‚baseline' 60,7% při přiřazování nejčastějšího valenčního rámce), viz citované práce J. Semeckého.

Další rozvíjení slovníku *VALLEX*

Výsledky uložené ve slovníku *VALLEX* prezentují valenci jako problém jak syntaktický (kombinatorický), tak lexikografický. Práce s bohatými daty ovšem ukazuje, že slovníkový přístup k zachycení valence přináší stále otevřené teoretické otázky, které vyžadují další podrobné lingvistické zkoumání.

Nejobtížnější problematikou při slovníkovém zpracování sloves zůstává vymezení jednotlivých významů. Při budování slovníku *VALLEX* se důraz klade na syntaktická kritéria, neboť jsou explicitnější než kritéria využívající hlubších rovin popisu. Zároveň je ovšem nepochybně nezbytné přihlížet k sémantice.

S problematikou vymezení významů též úzce souvisí (a do určité míry ji osvětluje) princip alternací, tedy schopnost sloves různým způsobem syntakticky strukturovat výpověď, viz oddíl 3. Koncept alternací je ovšem třeba dále rozpracovat – je potřeba jednak vytvořit gramatickou komponentu slovníku, která bude (na základě existujících studií) podrobně popisovat pravidla pro jednotlivé typy alternací, jednak u jednotlivých lexikálních jednotek ve slovníkové komponentě systematicky vyznačit možné alternace (dodejme, že technologicky a implementačně je alternační model slovníku připraven, viz [90]).

Obecněji pojaté alternace umožňují vzájemně provázat i slovesa s jejich prefigovanými deriváty, a alespoň na této úrovni provázat např. nedokonavá slovesa s jejich dokonavými prefigovanými protějšky.

Další problematikou, která zůstává ve středu zájmu, je možnost obohacení slovníku *VALLEX* o hlubší sémantické informace, zejména v návaznosti na existující datové zdroje. Zmiňme zde pilotní studie zaměřené na skupinu slovesa komunikace a výměny (‚communication' a ‚exchange', viz [31]), zkoumající možnost provázání těchto syntakticky i sémanticky heterogenních skupin sloves s vysoce oceňovanou sítí sémantických rámců *FrameNet*[20] (spojenou se jménem Ch. Fillmora), která je zpracovávána pro anglická slovesa, substantiva, adjektiva a adverbia [73]. Takovéto provázání by umožnilo i rozdělení českých sloves do jemnějších syntakticky i sémanticky homogenních tříd.

Literatura

[1] Anne Abeillé and Owen Rambow, editors. *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing.* Center for the Study of Language and Information, Stanford, 2000.

[2] K. Ajdukiewicz. Die syntaktische Konnexität. *Studia Philosophica*, I:1–27, 1935.

[3] Ondřej Bojar, Jiří Semecký, and Václava Benešová. Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *The Prague Bulletin of Mathematical Linguistics*, 83:5–17, 2005.

[20] `http://framenet.icsi.berkeley.edu/`

[4] Silvie Cinková. From PropBank to EngValLex. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2170–2175, Paris, 2006. ELRA.

[5] Silvie Cinková and Zdeněk Žabokrtský. Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns: Describing Event Structure in Support Verb Constructions. In Ferenc Kiefer, Gábor Kiss, and Júlia Pajzs, editors, *Proceedings of the 8th International Conference on Computational Lexicography COMPLEX*, pages 50–59, Budapest, 2005.

[6] Silvie Cinková and Zdeněk Žabokrtský. Treating Support Verb Constructions in a Lexicon: Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns. In Katrin Erk, Alissa Melinger, and Sabine Schulte im Walde, editors, *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 22–27, Saarbrücken, 2005.

[7] František Daneš. Větné členy obligatorní, potenciální a fakultativní. *Miscellanea Linguistica*, pages 131–138, 1971.

[8] František Daneš. *Věta a text*. Academia, Praha, 1985.

[9] František Daneš and Zdeněk Hlavsa. *Větné vzorce v češtině*. Academia, Praha, 1987. (aut. kolektiv: Jirsová, A. – Macháčková, E. – Prouzová, H. – Svozilová, N.).

[10] František Daneš, Miroslav Grepl, and Zdeněk Hlavsa, editors. *Mluvnice češtiny 3*. Academia, Praha, 1987.

[11] Charles J. Fillmore. The Case for Case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York, 1968.

[12] Charles J. Fillmore. Types of Lexical Information. In F. Kiefer, editor, *Studies in syntax and semantics*, pages 109–137. Kluwer Academic Publishers, New York, 1969.

[13] Miroslav Grepl and Petr Karlík. *Skladba češtiny*. Votobia, Olomouc, 1998.

[14] Jan Hajič. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Karolinum Press, Prague, 1998.

[15] Jan Hajič. Complex Corpus Annotation: The Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*, pages 54–73. Veda, Bratislava, 2006.

[16] Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-Coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume Vol. 9, pages 57–68, 2003.

[17] Eva Hajičová, Barbara H. Partee, and Petr Sgall. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer, Dordrecht, 1998.

[18] Eva Hajičová. K některým otázkám závislostní gramatiky. *Slovo a slovesnost*, 67(1):3–26, 2006.

[19] Bohuslav Havránek, editor. *Slovník spisovného jazyka českého*. Academia, Praha, 1964.

[20] Dana Hlaváčková. *Databáze slovesných valenčních rámců VerbaLex*. PhD thesis, Masarykova Universita, Brno, 2008.

[21] Dana Hlaváčková and Aleš Horák. Transformation of WordNet Czech Valency Frames into Augmented VALLEX-1.0 Format. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 310–313, Poznan, 2005. Wydawnictwo Poznańskie Sp. z o.o. with cooperation of Fundacja Uniwersytetu im. A. Mickiewicza.

[22] Dana Hlaváčková and Aleš Horák. VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In *Computer Treatment of Slavic and East European Languages*, pages 107–115, Bratislava, 2006. Slovenský národný korpus.

[23] Tomáš Holan, Vladislav Kuboň, Karel Oliva, and Martin Plátek. On Complexity of Word Order. *Les grammaires de dépendance – Traitement automatique des langues (TAL)*, 41(1):273–300, 2000.

[24] Aleš Horák. Verb Valency and Semantic Classification of Verbs. In P. Sojka, V. Matoušek, K. Pala, and I. Kopeček, editors, *Proceedings of Text, Speech and Dialog International Conference, TSD'98*, pages 61–66, Brno, 1998.

[25] Petr Jančar, František Mráz, Martin Plátek, and J. Vogel. On Monotonic Automata with a Restart Operation. *Journal of Automata, Languages and Combinatorics*, 4(4):287–311, 1999.

[26] Aravind Joshi. Tree Adjoining Grammars: How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions? In D. Dowty, editor, *Natural Language Processing*, pages 206–250. Cambridge University Press, Cambridge, 1985.

[27] Petr Karlík. Hypotéza modifikované valenční teorie. *Slovo a slovesnost*, 61:170–189, 2000.

[28] Petr Karlík, Marek Nekula, and Jana Pleskalová, editors. *Encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha, 2002.

[29] Václava Kettnerová. Czech Verbs of Communication with respect to the Types of Dependent Content Clauses. *The Prague Bulletin of Mathematical Linguistics*, 2008. (to appear).

[30] Václava Kettnerová. Konstrukce s rozpadem tématu a dikta v češtině. *Slovo a slovesnost*, 2008. (podáno).

[31] Václava Kettnerová, Markéta Lopatková, and Klára Hrstková. Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In *Proceedings of Text, Speech and Dialog International Conference, TSD 2008*, LNAI, pages 109-116, Berlin Heidelberg, 2008. Springer-Verlag.

[32] Karin Kipper, Benjamin Snyder, and Martha Palmer. Extending a Verb-lexicon Using a Semantically Annotated Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004, Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Paris, 2004. ELRA.

[33] Veronika Kolářová. Valence deverbálních substantiv: Některé specifické posuny v povrchových realizacích participantů. In Petr Karlík, editor, *Sborník konference Korpus jako zdroj dat o češtině*, pages 113–125, Brno, 2005.

[34] Veronika Kolářová. *Valence deverbativních substantiv v češtině.* PhD thesis, Univerzita Karlova v Praze, Praha, 2006.

[35] Beth C. Levin. *English Verb Classes and Alternations: A Preliminary Investigation.* The University of Chicago Press, Chicago and London, 1993.

[36] Beth C. Levin and Malka Rappaport Hovav. *Argument Realization.* Cambridge University Press, Cambridge, 2005.

[37] Markéta Lopatková. *Homonymie předložkových skupin a možnost jejich automatického zpracování.* PhD thesis, Univerzita Karlova v Praze, 2001.

[38] Markéta Lopatková. *O homonymii předložkových skupin v češtině (Co umí počítač?).* Nakladatelství Karolinum, Praha, 2003.

[39] Markéta Lopatková. Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics*, 79–80:37–60, 2003.

[40] Markéta Lopatková, Ondřej Bojar, Jiří Semecký, Václava Benešová, and Zdeněk Žabokrtský. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Proceedings of Text, Speech and Dialogue International Conference, TSD 2005*, volume 3658 of *LNAI*, pages 99–106, Berlin Heidelberg, 2005. Springer-Verlag.

[41] Markéta Lopatková and Jarmila Panevová. Valence vybraných skupin sloves (k některým slovesům dandi a recipiendi). In Z. Hladká and P. Karlík, editors, *Čeština – univerzália a specifika, Sborník konference ve Šlapanicích u Brna*, volume Vol. 5, pages 348–356. Nakladatelství Lidové noviny, Praha, 2004.

[42] Markéta Lopatková and Jarmila Panevová. Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*, pages 83–92. Veda, Bratislava, 2006.

[43] Markéta Lopatková and Jarmila Panevová. Valence vybraných sloves pohybu v češtině (antonyma, nebo synonyma?). In P. Piper, editor, *Sbornik Matice srpske za slavistiku*, volume 71-72/2007, pages 105–115, Novi Sad, 2007. Matica srpska.

[44] Markéta Lopatková, Martin Plátek, and Vladislav Kuboň. Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Proceedings of Text, Speech and Dialogue International Conference, TSD 2005*, volume 3658 of *LNAI*, pages 140–147, Berlin Heidelberg, 2005. Springer-Verlag.

[45] Markéta Lopatková, Martin Plátek, and Petr Sgall. Functional Generative Description, Restarting Automata and Analysis by Reduction. In F. Marušič and R. Žaucer, editors, *Studies in Formal Slavic Linguistics. Contributions from Formal Description of Slavic Languages 6.5.*, volume Vol. 19 of *Linguistik International*, pages 173–190. Peter Lang Publishing Group, Frankfurt am Main, 2008.

[46] Markéta Lopatková, Veronika Řezníčková, and Zdeněk Žabokrtský. Valency Lexicon for Czech: from Verbs to Nouns. In Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Proceedings of Text, Speech and Dialogue International Conference, TSD 2002*, volume 2448 of *LNAI*, pages 147–150, Berlin Heidelberg, 2002. Springer-Verlag.

[47] Markéta Lopatková, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha, 2008. (aut. kolektiv Skwarska, K. – Bejček, E. – Hrstková, K. – Nová, M. – Tichý, M.).

[48] Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarska, and Václava Benešová. VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report TR-2003-18, ÚFAL/CKL MFF UK, Prague, 2003.

[49] Markéta Lopatková and Zdeněk Žabokrtský. Testování konzistence a úplnosti valenčního slovníku českých sloves. In P. Vojtáš, editor, *Proceedings of ITAT 2003*, pages 73–82, Košice, 2003. University of P. J. Šafárik.

[50] Markéta Lopatková, Zdeněk Žabokrtský, and Karolína Skwarska. Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, volume Vol. 3, pages 1728–1733, Paris, 2006. ELRA.

[51] Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarska, and Václava Benešová. Tektogramaticky anotovaný valenční slovník českých sloves. Technical Report TR-2002-15, ÚFAL/CKL MFF UK, Praha, 2002.

[52] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[53] H. Messerschmidt, F. Mráz, F. Otto, and M. Plátek. Correctness Preservation and Complexity of Simple RL-Automata. In *Implementation and Application of Automata*, volume 4094 of *LNCS*, pages 162–172, Berlin Heidelberg, 2006. Springer-Verlag.

[54] František Mráz, Martin Plátek, and Friedrich Otto. Free Word-Order and Restarting Automata. In *Pre-proceedings of LATA 2007*, pages 425–436, Taragona, 2007. Universitat Rovira I Virgili.

[55] Friedrich Otto. Restarting Automata. In Z. Ésik, C. Martin-Vide, and V. Mitrana, editors, *Recent Advances in Formal Languages and Applications, Studies in Computational Intelligence*, volume Vol. 25, pages 269–303, Berlin, 2006. Springer-Verlag.

[56] Karel Pala and Pavel Ševeček. Valence českých sloves. In *Sborník prací FFBU*, pages 41–54, Brno, 1997.

[57] Martha Palmer, Dan Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005.

[58] Jarmila Panevová. On Verbal Frames in Functional Generative Description I-II. *The Prague Bulletin of Mathematical Linguistics*, 22-23:3–40,17–52, 1974-5.

[59] Jarmila Panevová. *Formy a funkce ve stavbě české věty.* Academia, Praha, 1980.

[60] Jarmila Panevová. Valency Frames and the Meaning of the Sentence. In Philip A. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. John Benjamins Publishing Company, Amsterdam, Philadelphia, 1994.

[61] Jarmila Panevová. Česká recipročí zájmena a slovesná valence. *Slovo a slovesnost*, 90:269–275, 1999.

[62] Jarmila Panevová. Poznámky k valenci podstatných jmen. In V. Karlík and Z. Hladká, editors, *Čeština – univerzália a specifika. Sborník konference ve Šlapanicích u Brna 17.-18. 11. 1998*, pages 173–180. Masarykova univerzita, Brno, 2000.

[63] Jarmila Panevová. K valenci substantiv (s ohledem na jejich derivaci). In P. Piper, editor, *Sbornik Matice srpske za slavistiku*, pages 29–36. Matica srpska, Novi Sad, 2003.

[64] Jarmila Panevová, Eva Benešová, and Petr Sgall. *Čas a modalita v češtině*, volume 34 of *Philologica Monographi*. Acta Universitatis Carolina, Praha, 1971.

[65] Vladimír Petkevič. A New Formal Specification of Underlying Structure. *Theoretical Linguistics*, 21(1):7–61, 1995.

[66] Martin Plátek. Composition of Translation with D-trees. In J. Horecký, editor, *Proceedings of the 9th International Conference on Computational Linguistics, COLING'82*, pages 313–318, Prague, 1982. Academia.

[67] Martin Plátek and Tomáš Holan. Závislostní a složkové modelování syntaxe jazyku. In D. Obdržálek and J. Štanclová, editors, *Malý informatický seminář, MIS 2004*, pages 115–150, Praha, 2004. Matfyzpress.

[68] Martin Plátek and Markéta Lopatková. Funkční generativní popis a formální teorie překladů. In P. Vojtáš, editor, *Proceedings of ITAT 2007*, pages 3–14, Košice, 2007. University of P. J. Šafárik.

[69] Martin Plátek, František Mráz, Friedrich Otto, and Markéta Lopatková. O roztržitosti a volnosti slovosledu pomocí restartovacích automatu. In P. Vojtáš, editor, *Proceedings of ITAT 2005*, pages 145–156, Košice, 2005. University of P. J. Šafárik.

[70] Martin Plátek and Friedrich Otto. A Two-Dimensional Taxonomy of Proper Languages of Lexicalized FRR-Automata. In *Pre-proceedings of LATA 2008*, Taragona, 2008. Universitat Rovira I Virgili.

[71] Martin Plátek, Jiří Sgall, and Petr Sgall. A Dependency Base for a Linguistic Description. In Petr Sgall, editor, *Contribution to Functional Syntax, Semantics and Language Comprehension*, volume 16 of *Linguistic and Literary Studies in Eastern Europe*, pages 63–97. Academia, Prague, 1985.

[72] Martin Plátek and Petr Sgall. A Scale of Context-Sensitive Languages: Applications to Natural Language. *Information and Control*, 38(1):1–20, 1978.

[73] Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. *FrameNet II: Extended Theory and Practice*. University of California, Berkeley, 2006. http://framenet.icsi.berkeley.edu/book/book.html.

[74] Anoop Sarkar and Daniel Zeman. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics, COLING 2000*, pages 691–697, Saarbrücken, Germany, 2000.

[75] Jiří Semecký. *Verb Valency Frames Disambiguation*. PhD thesis, Charles University in Prague, Prague, 2007.

[76] Jiří Semecký and Petr Podveský. Extensive Study on Automatic Verb Sense Disambiguation in Czech. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *Proceedings of Text, Speech and Dialog International Conference, TSD 2006*, volume 4188 of *LNAI*, pages 237–244, Berlin Heidelberg, 2006. Springer-Verlag.

[77] Petr Sgall. *Generativní popis jazyka a česká deklinace*. Academia, Praha, 1967.

[78] Petr Sgall. Teorie valence a její formální zpracování. *Slovo a slovesnost*, 59:15–29, 1998.

[79] Petr Sgall. Valence jako jádro jazykového systému. *Slovo a slovesnost*, 67:163–178, 2006.

[80] Petr Sgall, Allevtina Bémová, Jan Borota, Eva Hajičová, Ivana Hajičová, Petr Jirku, Jarmila Panevová, Martin Plátek, and Jarka Vrbová. *Úvod do syntaxe a sémantiky*. Academia, Praha, 1986.

[81] Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

[82] Petr Sgall, Ladislav Nebeský, Alla Goralčíková, and Eva Hajičová. *A Functional Approach to Syntax in Generative Description of Language*. American Elsevier Publishing Company, Inc., New York, 1969.

[83] Petr Sgall, Jarmila Panevová, and Eva Hajičová. Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38, Boston, 2004. Association for Computational Linguistics.

[84] Hana Skoumalová. *Czech Syntactic Lexicon*. PhD thesis, Charles University in Prague, 2001.

[85] Hana Skoumalová, Markéta Straňáková-Lopatková, and Zdeněk Žabokrtský. Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In V. Matoušek, P. Mautner, R. Mouček, and K. Taušer), editors, *Proceedings of Text, Speech and Dialog International Conference, TSD 2001*, volume 2166 of *LNAI*, pages 142–149, Berlin Heidelberg, 2001. Springer-Verlag.

[86] Markéta Straňáková-Lopatková and Zdeněk Žabokrtský. Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In Manuel González Rodríguez and Carmen Paz Suárez Araujo, editors, *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, volume Vol. 3, pages 949–956, Paris, 2002. ELRA.

[87] Naďa Svozilová, Hana Prouzová, and Anna Jirsová. *Slovesa pro praxi*. Academia, Praha, 1997.

[88] Lucien Tesnière. *Eléments de syntaxe structurale*. Librairie C. Klincksieck, Paris, 1959.

[89] Zdeňka Urešová. The Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*, pages 93–112. Veda, Bratislava, 2006.

[90] Zdeněk Žabokrtský. *Valency Lexicon of Czech Verbs*. PhD thesis, Charles University in Prague, Prague, 2005.

[91] Zdeněk Žabokrtský and Markéta Lopatková. Valency Frames of Czech Verbs in VALLEX 1.0. In Adam Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 70–77, Boston, 2004. Association for Computational Linguistics.

[92] Daniel Zeman. *Parsing with a Statistical Dependency Model*. PhD thesis, Charles University in Prague, Prague, 2004.

[93] Daniel Zeman and Anoop Sarkar. Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC 2000*, volume Vol. 1, pages 227–233, Athene, Greece, 2000.

[94] George Kingsley Zipf. *Psycho-Biology of Languages*. Houghton-Mifflin, Boston, 1935. (2nd edition MIT Press, 1965).