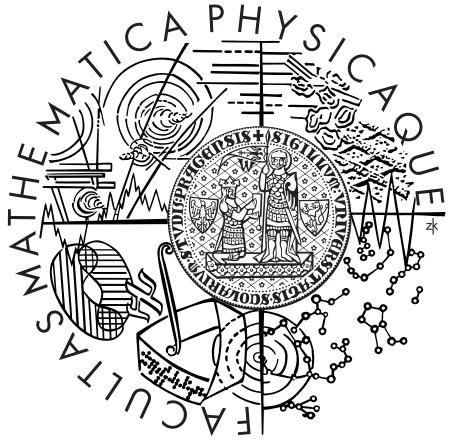
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Disertační práce



AUTOMATICKÉ PROPOJOVÁNÍ LEXIKOGRAFICKÝCH ZDROJŮ A KORPUSOVÝCH DAT

Eduard Bejček

Praha, 2015

Disertační práce

Eduard Bejček

Vedoucí disertační práce:
doc. RNDr. Markéta Lopatková, Ph.D.

Automatické propojování lexikografických zdrojů a korpusových dat

Studijní program: Informatika
Studijní obor: Matematická lingvistika

Vysázeno LaTeXem.
Praha, 2015

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

Prohlašuji, že jsem tuto disertační práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 19. června 2015 Eduard Bejček

Název práce:	Automatické propojování lexikografických zdrojů a korpusových dat
Autor:	Eduard Bejček
Ústav:	Ústav formální a aplikované lingvistiky
Vedoucí disertační práce:	doc. RNDr. Markéta Lopatková, Ph.D., Ústav formální a aplikované lingvistiky
Klíčová slova:	prolinkování; slovník; valence; víceslovné výrazy

Abstrakt: Spolu se vznikem stále dalších jazykových zdrojů – slovníků, lexikálních databází, korpusů, treebanků – roste i potřeba jejich účinného propojování, které by umožnilo snadné využití veškerých shromážděných vlastností a informací. V tomto ohledu je také aktuální téma univerzálních lexikografických formátů.

Tato práce zkoumá metody automatického propojování jazykových dat. Představíme zde systém na propojování slovníků, jakými jsou například VALLEX, PDT-Vallex, FrameNet, nebo SemLex, které poskytují syntaktickou informaci o svých heslech. Systém je automatický, umožňuje tudíž opakovanou aplikaci na novější verze vyvíjejících se jazykových zdrojů. Na základě syntaktické informace obsažené ve slovníku víceslovných výrazů SemLex navrhujeme metodu vyhledávající tyto výrazy v automaticky anotovaném textu.

Praktickým výstupem potvrzujícím úspěšnost použitých metod je mj. propojení slovníků VALLEX a PDT-Vallex vedoucí k doplnění desítek tisíc anotovaných vět z treebanků PDT a PCEDT do VALLEXu.

Title: Automatic linking of lexicographic sources and corpus data
Author: Eduard Bejček
Department: Institute of Formal and Applied Linguistics
Supervisor: doc. RNDr. Markéta Lopatková, Ph.D.,
Institute of Formal and Applied Linguistics
Keywords: linking; lexicon; valency; multiword expressions

Abstract: Along with the increasing development of language resources – i.e., new lexicons, lexical databases, corpora, treebanks – the need for their efficient interlinking is growing. With such a linking, one can easily benefit from all their properties and information. Considering the convergence of resources, universal lexicographic formats are frequently discussed.

In the present thesis, we investigate and analyse methods of interlinking language resources automatically. We introduce a system for interlinking lexicons (such as VALLEX, PDT-Vallex, FrameNet or SemLex) that offer information on syntactic properties of their entries. The system is automated and can be used repeatedly with newer versions of lexicons under development. We also design a method for identification of multiword expressions in a parsed text based on syntactic information from the SemLex lexicon.

An output that verifies feasibility of the used methods is, among others, the mapping between the VALLEX and the PDT-Vallex lexicons, resulting in tens of thousands of annotated treebank sentences from the PDT and the PCEDT treebanks added into VALLEX.

Poděkování

Rád bych vyjádřil vděčnost své školitelce Markétě Lopatkové, která mě během celé práce vedla i povzbuzovala, učila i motivovala. Bez jejích vlídných doporučení by tento text nezačal vznikat dodnes a nemohl bych jí tedy ani poděkovat za neúnavné a pečlivé čtení postupně vznikající práce.

Děkuji všem kolegům z Ústavu formální a aplikované lingvistiky za výborné pracovní, stejně jako skvělé neformální prostředí. Děkuji i ústavu samotnému za možnost prožít zde třináct krásných studentských let. Za konzultace a četné rady chci poděkovat zejména Pavlu Straňákovi, Vendule Kettnerové, prof. Jarmile Panevové, prof. Evě Hajičové a všem dalším, se kterými jsem měl tu radost a čest diskutovat.

Zmíněným i nezmíněným kolegům děkuji také za ohleduplnost, se kterou okolo mě v posledních měsících chodili po špičkách a neodvažovali se mě požádat o spolupráci na čemkoli; namísto toho mi dodávali povzbuzení do další práce.

Mimořádné poděkování patří rodičům a celé rodině za podporu během mého studia. Ačkoli to je nyní u konce, vím, že v nich mám oporu i nadále.

Na závěr chci poděkovat všem přátelům, kteří věřili, že práci dopíšu, často víc než já sám.

Tato práce vznikla mj. za podpory grantu GA P406/12/0557 Grantové agentury České republiky a grantu COST CZ LD14117 Ministerstva školství, mládeže a tělovýchovy České republiky. Výzkum publikovaný v této práci využívá jazykové zdroje vyvinuté, uložené a distribuované projektem LINDAT/CLARIN Ministerstva školství, mládeže a tělovýchovy České republiky (projekt LM2010013).

Obsah

1

Úvod

Počítačová lingvistika od svého vzniku pracuje s elektronickými slovníky, posléze i s textovými korpusy, bez nichž je dnes již seriozní lingvistická práce nepředstavitelná. Hodnota korpusu se zvyšuje, je-li spojen se slovníkem: k označeným jednotkám textu je pak možné získat ze slovníku dodatečné informace a naopak informace ze slovníku jsou mnohem použitelnější, jsou-li exemplifikovány skutečnými výskyty, často v řádu tisíců.

Výroba takových jazykových zdrojů je pracná, a tudíž finančně nákladná i časově náročná. Sílí proto poptávka po vytváření těchto zdrojů automaticky. Jedním ze způsobů, jak se tomu lze přiblížit, je využít existující, pracně vyrobené jazykové zdroje k tvorbě zdrojů nových, nebo k jejich obohacování.

Tato práce pojednává o automatickém propojení několika valenčních slovníků, o propojení valenčního slovníku s korpusem, propojení slovníku víceslovných výrazů s korpusem a v závěru se lehce dotkne vnitřních vztahů ve slovníku a propojení slovníkových položek mezi sebou.

Motivace

Domníváme se, že ani sebevíce propracovaná gramatika přirozeného jazyka nedokáže zachytit všechny jeho aspekty – bez slovníku. Slovník (či lexikální databáze) je místo, kde se k základním elementům jazyka, typicky ke slovům, přistupuje jednotlivě a každý popisovaný element zde má prostor pro uvedení svých vlastních, jedinečných vlastností: v jakých kontextech se používá, jak se kombinuje s jinými elementy, jakých může nabývat tvarů a variant, jaký má význam, jaký vztah má k jiným elementům…

Diskuse o významu gramatiky a slovníku je ovšem letitá (a její teoretické řešení není pro naši práci podstatné). Krátký historický přehled přináší Panevová a Ševčíková (2014). Za zástupce těch, kteří se přiklonili k nadřazené roli gramatiky, jmenují Bloomfielda (1933) a přinejmenším starší verze Chomského transformační gramatiky (Chomsky, 1957) (později zde význam slovníku poněkud vzrostl).

Na druhé straně propagující slovníky zmiňují autorky teorii Smysl–Text (Meaning-Text theory, Mel'čuk, 1988), která z rozsáhlého a strukturovaného slovníku

čerpá, dále kategoriální gramatiky (Ajdukiewicz, 1935) či lexikalizovanou „tree adjoining grammar“ (Joshi a Schabes, 1991).

Z pohledu této práce je relevantní hledisko, které formuloval Chomsky (1970), tedy že odpověď na otázku, zda většina potřebné informace náleží slovníku či gramatice, není obsažena v jazyce samém, ale je věcí empirickou („entirely an empirical issue“).

Analýzu přínosu slovníků k různým aplikacím zpracování přirozeného jazyka přináší například Bojar (2009). Nachází pochopitelně případy potvrzující oba přístupy: práce, kde slovníky dané úloze nepomáhají, i opačné případy, ve kterých zapojení slovníků do systému zlepší výsledky. Příkladem úspěšného zapojení slovníku je třeba aplikace pro vyhledávání informací, kde je slovník využíván k rozšíření vyhledávacího dotazu (zvýšení „recall“), dále zlepšení přesnosti syntaktické analýzy zapojením subkategorizace (která často je, ale ne vždy musí být založena na explicitně dopředu připraveném slovníku) nebo některé strojové překladové systémy.

Využití slovníku zlepšuje mnoho úloh počítačového zpracování jazyka, pro některé je pak slovník zcela nezbytný. Jiné přístupy se bez něj však obejdou.

Vznikalo a vzniká proto velké množství slovníků. Zachycují různé aspekty jazyka, vychází z různých teoretických základů, vznikají v různých formátech a pro různé přirozené jazyky. Výroba každého jednoho takového slovníku je velice nákladná záležitost. Rovněž použití slovníku pro anotaci většího množství textu, které je pro mnoho aplikací klíčové a pro slovník výhodné kvůli validaci, je pracné a drahé. Pro mnoho jazyků již přesto existuje vedle sebe více odlišných slovníků, často propojených s některými textovými korpusy. Obvykle žádný z nich neobsahuje kompletní informaci. Je proto logické, že je v posledních letech viditelný trend směřující k jejich propojování, což zvyšuje pokrytí a rozšiřuje spektrum informací ve slovníku obsažených. Také je patrná snaha o propojení slovníků napříč jazyky, neboť ruční výroba určitého slovníku v novém jazyce vyžaduje stejné množství manuální práce a peněz jako výroba slovníku původního.

Celá řada projektů se již zabývala propojováním slovníků. Mezi jinými jmenujme pro angličtinu rozsáhlý projekt SemLink představený v sekci 4.8 (Loper, Yi a Palmer, 2007)[1] a návazné projekty WordFrameNet (Laparra, Rigau a Cuadros, 2010) a Predicate Matrix (de Lacalle, Laparra a Rigau, 2014), dále spojení WordNetu a FrameNetu s názvem MapNet (Tonelli a Pighin, 2009) nebo pro češtinu propojení VerbaLexu s FrameNetem (Materna a Pala, 2010). Jednomu z našich úkolů je blízké propojování dvou španělských slovníků popisujících „sub-

[1] Titíž autoři také rovnou uvádějí, že spojení s VerbNetem zvýšilo úspěšnost přiřazování sémantických rolí z PropBanku (semantic role labeling).

kategorizační rámce“, tedy vlastně (povrchovou) valenci sloves (Necsulescu et al., 2011).

Témata

Jazykové zdroje, které chápeme jako (i) slovníky a lexikální databáze a (ii) textová data, anotované korpusy a treebanky, je možné užitečně propojovat vícerými způsoby. Pro přehlednost úlohy rozdělíme do pěti skupin. Třem z nich se budeme více věnovat.

1) slovník–slovník Propojování dvou zdrojů typu (i). Typicky se jedná o propojení významových jednotek slovníků: např. synsetů ve WordNetu, lexikálních jednotek ve VALLEXu či sémantických rámců ve FrameNetu. Můžeme však propojovat i specifičtější jednotky, nebo naopak může být výhodné i prosté propojení lemmat. Přinese větší pokrytí (propojované slovníky nemají stejný repertoár slovníkových hesel) a rozšíření informací (propojované slovníky nezachycují stejné údaje o svých položkách).

2) slovník–text Propojování (i) a (ii). Ruční (ale i automatické) propojování je obvykle nazýváno anotací. Slouží k ověření použitelnosti slovníku a k přiřazení dokladů k jednotlivým heslům, stejně jako k obohacení korpusu a desambiguaci tokenů.

3) uvnitř slovníku Propojování vybraných jednotek mezi sebou (tedy zachycování vnitřních vztahů mezi různými jednotkami téhož slovníku) v jednom zdroji typu (i) pomáhá teoretickému zpracování jazyka, hledání souvislostí ve slovnících a jevů, jejichž pravidelnost vyplyne až porovnáním. Napomáhá také ekonomickému popisu, díky němuž je snazší udržet slovník v konzistentním stavu.[2]

V této práci se věnujeme zejména skupině (1) a (2), okrajově pak skupině (3).

Cíle práce

Naším cílem je zkoumat metody pro propojování slovníků, resp. jejich lexikálních jednotek, navržené metody implementovat a otestovat. Výsledný systém bude automatický a tudíž pro podobné zdroje použitelný opakovaně, navíc s využitím

[2] Zcela mimo rámec této práce leží zbylé dvě skupiny: propojování jednotek **4) uvnitř textu** a propojování **5) text-text**. Do čtvrté skupiny můžeme zařadit spojování slov v (ii) do skupin jako jsou např. víceslovné výrazy nebo klauze, ale také např. koreferenční vztahy mezi jednotlivými vzdálenými slovy, nebo diskurzní vztahy mezi většími celky slov. Poslední pátá skupina je tvořena paralelními (vícejazyčnými) korpusy.

případných manuálních oprav výsledků z minulosti. Zejména se zaměříme na slovníky, kde jsou lexikální jednotky popisovány na základě syntaktických vlastností, a to vlastností náležejících hloubkové i povrchové rovině jazyka. Tohoto syntaktického, jinak též valenčního popisu využijeme pro porovnání slovníkových hesel a spojení těch, která si odpovídají.

Dále předložíme metodu pro vyhledávání slovníkových položek delších než jedno slovo, víceslovných výrazů. Tato slovníková hesla budeme v textu vyhledávat na základě syntaktických vztahů mezi dílčími slovy, tedy s využitím vyšších rovin jazykového popisu.

Při práci s různými lexikografickými zdroji je důležitým aspektem formát každého z nich, neboť musí být snadné různě kódovanou informaci obsaženou v obou převést do formy, která umožňuje její porovnání. K tomuto účelu se nabízí využít některé z univerzálních formátů a lexikografických standardů, čemuž však musí předcházet analýza, zda je takový formát skutečně přínosným řešením.

Propojovací metody představíme na konkrétních případech dvojic jazykových dat. Konkrétní cíle se tak rozpadají na následující úkoly. Zaměříme se zejména na zpracování slovníků VALLEX, PDT-Vallex a SemLex, které budou popsány dále (sekce 4.2, 4.3 a 4.9). Na nich ověříme navržené postupy a otestujeme různé konkrétní způsoby propojování. Nejprve tedy představíme řadu relevantních valenčních slovníků a lexikálních databází pro češtinu a angličtinu; zvláštní zřetel věnujeme jejich formátům. Poté přistoupíme k řadě experimentů, jejichž cílem je propojování různých dvojic jazykových zdrojů: slovníků i korpusů.

Konkrétní cíle této práce jsou potom zejména následující:

- Propojit české valenční slovníky VALLEX a PDT-Vallex na úrovni lexikálních jednotek tam, kde si odpovídají, kde obě vyjadřují „stejný“ význam zapsaný „stejným“ valenčním rámcem.
- Umožnit v budoucnu opakovaně propojovat tyto dva slovníky, neboť jejich sjednocení se neplánuje a budou se nadále (odděleně) vyvíjet.
- Obohatit vzorek sloves z VALLEXu o příkladové věty z projektu VALEVAL.
- Propojit VALLEX s anotovanými větami v PDT a PCEDT.
- Navrhnout metodu pro vyhledání víceslovných výrazů ze slovníku SemLex v libovolném textu.
- Seznámení s formáty běžně používaných slovníků včetně univerzálních slovníkových formátů a doporučovaných standardů standardizačních iniciativ s ohledem na jejich možné uplatnění v některém ze shora uvedených úkolů.

2

Přehled částí práce

Po úvodních kapitolách 1–2 představíme v **první části** korpusová data a lexikografické zdroje, které se bezprostředněji týkají této práce, zejména všechny, se kterými budeme následně pracovat.

V kapitole 3 Korpusy textové i syntaktické postupně popíšeme Český národní korpus, anglický syntaktický složkový korpus Penn Treebank a dále české závislostní korpusy PDT s pouze českými texty (verzi 2.0 i novinky pozdějších verzí) a PCEDT s paralelní anotací pro češtinu a angličtinu. O všech zmíněných korpusech budeme hovořit ve druhé části práce.

Kapitola 4 Slovníky a jejich formáty se zaměřuje zejména na valenční slovníky, začíná proto sekcí o valenci. Následují české valenční slovníky VALLEX, PDT-Vallex a přehled několika dalších českých slovníků, které však nevyužíváme. Poté přejdeme k lexikálním databázím. Popis začneme anglickým a českým WordNetem, což je zároveň první ukázka provázaných lexikálních zdrojů. Zůstaneme u angličtiny a představíme databázi FrameNet, ale i ostatní lexikální databáze, které jsou všechny navzájem propojené v rámci projektu SemLink. Poslední představený slovník bude český slovník víceslovných výrazů SemLex. Každá sekce končí popisem datového formátu, v jakém je slovník k dispozici, s výjimkou dvou hlavních slovníku VALLEX a PDT-Vallex, jejichž formátům jsou věnovány samostatné sekce. Kvůli další práci s těmito slovníky formáty porovnáme a navrhneme jeden společný formát pro oba slovníky. Na závěr vysvětlíme, proč byl tento postup výhodnější než použití některého ze standardních, univerzálních formátů. Slovníky VALLEX, PDT-Vallex, SemLex a částečně FrameNet budeme využívat v druhé části práce.

Druhá část práce nás provede projekty, jejichž cílem je propojování nejrůznějších dvojic jazykových zdrojů. Půjde zejména o projekty, na nichž se podílel autor této práce, ale pro úplnost zmíníme okrajově i některé další projekty. Druhá část je rozdělena na tři kapitoly na základě typu provazovaných zdrojů: propojujeme zde dvojice slovníků, dvojice slovník–korpus nebo různé jednotlivé prvky uvnitř jednoho slovníku mezi sebou.

V kapitole 5 Propojování slovníků mezi sebou popíšeme projekt autora práce s názvem Vallink, který automaticky propojil valenční slovníky VALLEX a PDT-Vallex na úrovni lexikálních jednotek. Popis nutné přípravy v podobě návrhu společného formátu jsme podali v první části, zde popíšeme samotný postup pravidlového porovnávání lexikálních jednotek a výstup vyhodnotíme oproti baseline na 200 testovacích slovesech (gold datech). Dále popíšeme zjednodušenou variantu projektu Vallink nula, která nám umožnila propojit starší slovník VALLEX 1.0 s novým VALLEXem 2.5. Popíšeme též postup, kterým automaticky propojujeme VALLEX a FrameNet (a PDT-Vallex) v probíhajícím projektu Vallink II.

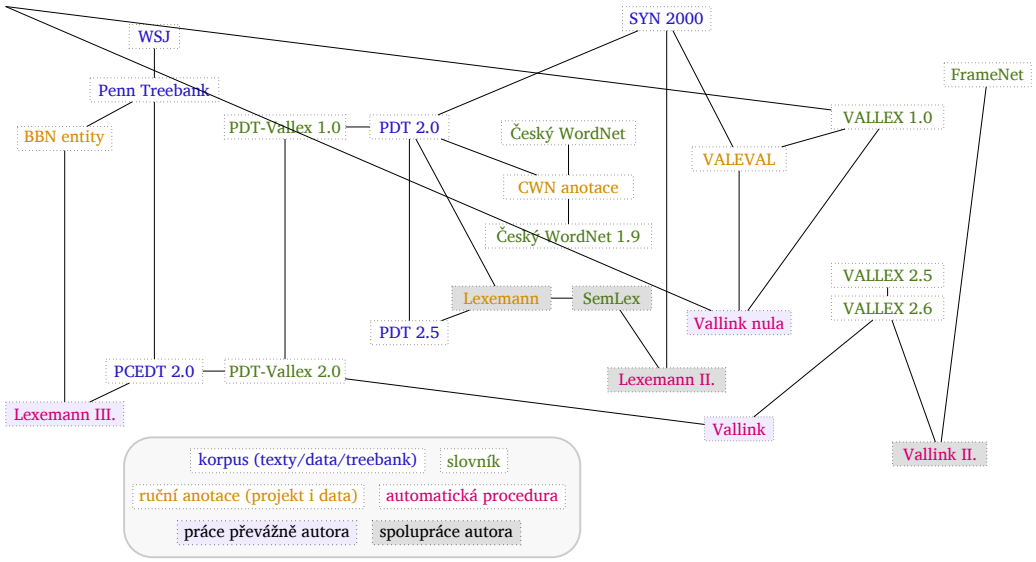
V kapitole 6 Propojování slovníku s textem přejdeme ke kombinaci slovníku a korpusu. Po uvedení do problematiky „víceslovných lexií“ a představení staršího projektu Lexemann popisujeme automatickou identifikaci slovníkových víceslovných lexií v anotovaném textu (zde v obrázku 2.1 označeno jako Lexemann II). Tím dochází k propojení textu a slovníku, kdy ke slovníkovému heslu přibývají doklady z korpusu. Na to navážeme, když stejnou metodu použijeme k automatické identifikaci víceslovných (pojmenovaných) entit v anotovaném textu (označeno Lexemann III.). Identifikaci víceslovných entit využíváme spíše pro kontrolu konzistence anotací. Poslední dvě kratší sekce staví na výsledcích propojení dvojic slovníků: díky propojení dvou verzí VALLEXu získáváme pro novou verzi navázání vybraných lexikálních jednotek slovníku na výskyty v textu vybrané z ČNK a díky propojení VALLEXu a PDT-Vallexu získáváme pro více než čtvrtinu lemmat navázání lexikálních jednotek na výskyty v syntakticky anotovaném PDT.

Sekce 7 dá do souvislosti rozšiřování použitelnosti dat propojováním informací napříč jazykovými zdroji s propojováním informací uvnitř elektronických slovníků. Uvedeme některé příklady současného stavu a představíme projekt provazování lexikálních jednotek mezi sebou ve slovníku VALLEX (relací alternace a svým způsobem i diateze). Na závěr se vrátíme ke slovníku víceslovných výrazů SemLex, pro nějž popíšeme plány na provázání dvou typů vnitřních vztahů: souvisejících a zanořených výrazů.

Pro snazší orientaci v projektech a zdrojích přikládáme ještě tabulku a obrázek, které pokrývají všechny popisované a související projekty z kapitol 5 a 6, dále specifikují, které části jsou dílem autora (tabulka obsahuje odkazy na příslušné části této práce).

typ	co s čím	projekt	spoluautor	odkaz
S–S	VALLEX + PDT-Vallex	Vallink	—	5.1
S–T	VALLEX + PDT	*důsledek* ↘	—	6.4
S–S	VALLEX + FrameNet	Vallink II.	V. Kettnerová	5.3
S–S	PDT-Vallex + FrameNet	*důsledek* ↘	—	5.4
S–S	VALLEX + VALLEX	Vallink nula	—	5.2
S–T	VALLEX + ČNK	*důsledek* ↘	—	6.3
S–T	SemLex + PDT	Lexemann	P. Straňák	Straňák (2010)
S–T	SemLex + ČNK	Lexemann II.	P. Straňák a P. Pecina	6.1
S–T	BBN + PCEDT	Lexemann III.	— (P. Straňák)	6.2

Tabulka 2.1: Přehled projektů, o nichž pojednává tato práce. První sloupeček představuje typ propojování: slovníků (S) či textových dat (T).



Obrázek 2.1: Přehled jazykových zdrojů použitých v této práci a projektů, které zde popisujeme. Graf je řazen shora dolů přibližně chronologicky. Projekty, které jsou pro tuto práci relevantní, jsou v případě účasti autora práce podbarveny šedě či fialově.

Část I

Jazykové zdroje a jejich formáty

I. JAZYKOVÉ ZDROJE A JEJICH FORMÁTY

V první části představíme řadu lexikálních zdrojů. Některé z nich budeme používat ve zbytku této práce, jiné však popíšeme pouze za účelem doplnění širšího obrázku o spektru těchto zdrojů.

Kratší kapitola 3 je věnována korpusům a treebankům, následující kapitola 4 se bude týkat slovníků, zejména valenčních slovníků nebo s nimi příbuzných lexikálních databází.

3

Korpusy textové i syntaktické

Jazykový korpus je databáze sesbíraná z reálných textů (napsaných za jiným účelem než tvorby korpusu). Pokud se podaří korpus sestavit tak, aby v něm byly vyváženě zastoupeny různé žánry, může sloužit jako referenční zdroj pro testování jazykových hypotéz, pro tvorbu slovníků, frekvenčních slovníků, pro ověřování gramatických pravidel, pro zjišťování pravidelností, hledání kolokací nebo třeba pro tvorbu statistického jazykového modelu.

Častější jsou korpusy synchronní (obsahující současný jazyk – oproti diachronním) a jednojazyčné (oproti vícejazyčným, paralelním), neboť je také snazší je sestavit. Podobně největší korpusy (například sesbírané z webu) nejsou vyvážené.

Moderní korpusová lingvistika samozřejmě pracuje s elektronickými korpusy s efektivním vyhledáváním – nicméně například Kancelář Slovníku jazyka českého (od roku 1946 Ústav pro jazyk český při Akademii věd) pořizovala excerpty z české literatury do lístkové Kartotéky lexikálního archivu již od roku 1911 (a to i zpětně, pro texty od roku 1770).

S nástupem elektronických archivů bylo stále snazší (a také stále potřebnější) dodat k obyčejnému korpusu ručně další informace (tedy dodatečnou anotaci), podle kterých by bylo možné vyhledávat. Vznikají tak korpusy stále složitější a bohatší. Nejjednodušší značkování spočívá v připojení lemmatu či morfologické informace (typicky obojího) ke každému slovu.

Stále ještě se pohybujeme na rovině *textových korpusů*, jakým je například Český národní korpus (sekce 3.1). Se složitější anotací syntaktické struktury vznikají tzv. *treebanky*, neboli *syntaktické korpusy*. Mezi ně patří například korpus se složkovou syntaxí Penn Treebank (sekce 3.2) a korpus se závislostní syntaxí Pražský závislostní treebank (sekce 3.3 a 3.4).

Na ještě vyšší rovině anotace se pohybují tzv. *sémantické treebanky*, mezi které lze mj. zařadit FrameNet, o kterém píšeme v sekci 4.6 mezi slovníky.

V poslední sekci 3.5 této kapitoly budeme mluvit o paralelním česko-anglickém treebanku PCEDT.

3.1 ČNK

Ústav Českého národního korpusu[1] (ÚČNK) spravuje velkou řadu různorodých korpusů, většina z nich je pro češtinu. Nás bude nejvíce zajímat synchronní řada současných textů SYN (přes dvě miliardy slov), do níž spadají vedle publicistických korpusů také žánrově vyvážené korpusy, mj. SYN 2000[2] čítající 100 milionů slov. Všechny korpusy řady SYN jsou automaticky označkované: každé slovo má přiřazeno lemma a poziční morfologickou značku v Pražském pozičním tagsetu.[3]

Mezi další korpusy ÚČNK patří například korpus soukromé korespondence, korpus školních písemných prací, reprezentativní korpus neformální mluvené češtiny, brněnský mluvený korpus, diachronní korpus, webový korpus němčiny, korpusy lužických srbštin či paralelní korpusy.

3.2 Penn Treebank

Penn Treebank[4] byl prvním velkým syntakticky anotovaným korpusem, neboli treebankem.

Obsahuje milion slov získaných z Wall Street Journalu (WSJ) a dále některé kratší texty (přepisy mluvené řeči ATIS-3 a Switchboard). Syntaxe všech vět je ručně zpracována do podoby složkových stromů, jejichž ukázku vidíme na obrázku 3.1.

Na obrázku vidíme neterminály v hranatých rámečcích a terminály značené kroužkem, všechny v dolní řadě. Neterminály jsou ohodnoceny typem fráze – např. S-TPC (topikalizovaná klauze), NP (jmenná fráze), VP (slovesná fráze) apod. – terminály mají kromě slovního tvaru přiřazen slovní druh – VBP (sloveso, které není ve 3. osobě singuláru présentu), DT (člen), CC (koordinační spojka), JJ (adjektivum) apod. Vidíme také jeden výskyt tzv. *trace*[5] (*T*), která obsazuje místo předmětu slovesa *say* a referenční šipkou odkazuje na celou vedlejší větu, která ve skutečnosti onen předmět tvoří.

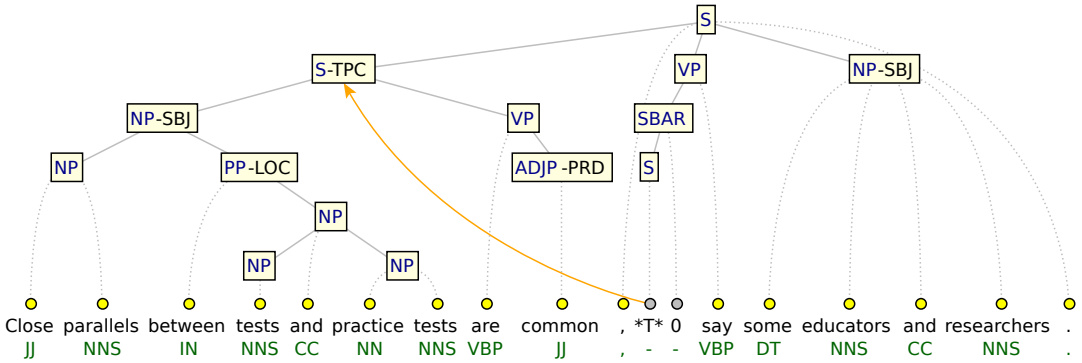
[1] `https://ucnk.ff.cuni.cz`

[2] *Český národní korpus – SYN2000.* Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: `http://www.korpus.cz`.

[3] `https://ucnk.ff.cuni.cz/doc/popis_znacek.pdf`

[4] `https://www.cis.upenn.edu/~treebank`, `https://catalog.ldc.upenn.edu/LDC99T42`

[5] *Traces*, česky také *stopy*, je pomocný koncept transformačních gramatik, který řeší změny slovosledu (například v tázacích větách nebo při pasivizaci). Je to prázdná kategorie, která ale obsazuje potřebné místo v syntaktické struktuře.



Obrázek 3.1: Ukázka věty ve složkovém stromu anotovaném v rámci Penn Treebanku:
„Close parallels between tests and practice tests are common, some educators and researchers say.“

Penn Treebank v této práci zmíníme dvakrát: posloužil jako materiál pro anotaci PropBanku (který jen představíme v sekci 4.7.1) a také pro anotaci tzv. BBN entit (jejíž konzistenci spolu s konzistencí PCEDT otestujeme v sekci 6.2).

3.3 PDT 2.0

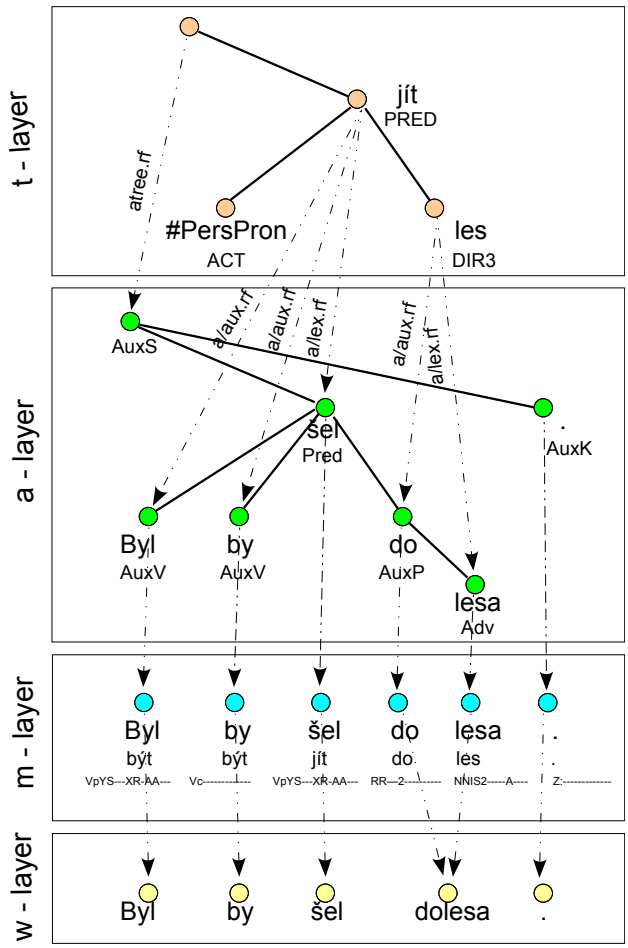
Pražský závislostní treebank 2.0 (Prague Dependency Treebank, PDT)[6] je jedním z prvních hloubkově anotovaných treebanků na světě. Po pětileté práci bylo na přelomu let 2000/2001 vydáno PDT 1.0 (Hajič et al., 2001), které obsahovalo pouze morfologické a povrchově syntaktické informace. Po dalších 6 letech bylo dosaženo stavu, který byl v hrubých rysech shodný s dnešním. PDT 2.0 obsahuje anotace na morfologické, analytické a hloubkově syntaktické rovině.

PDT je postaven na teorii FGP (zejména Sgall, 1967; Sgall, Hajičová a Panevová, 1986) a vznikal mj. jako způsob, jak ji ověřit v praxi, na skutečných větách. Téměř všechny části anotace byly prováděny ručně, či přinejmenším ručně kontrolovány. Během anotací vznikly objemné manuály (Mikulová et al., 2006; Hajič et al., 2004; Hana et al., 2005), které do detailu popisují, jak se má v kterých případech postupovat.

Výsledkem jsou novinové texty z ČNK v rozsahu 2 000 000 slov anotované na čtyřech rovinách jazykového popisu. Pro zběžné přiblížení použijeme tradiční

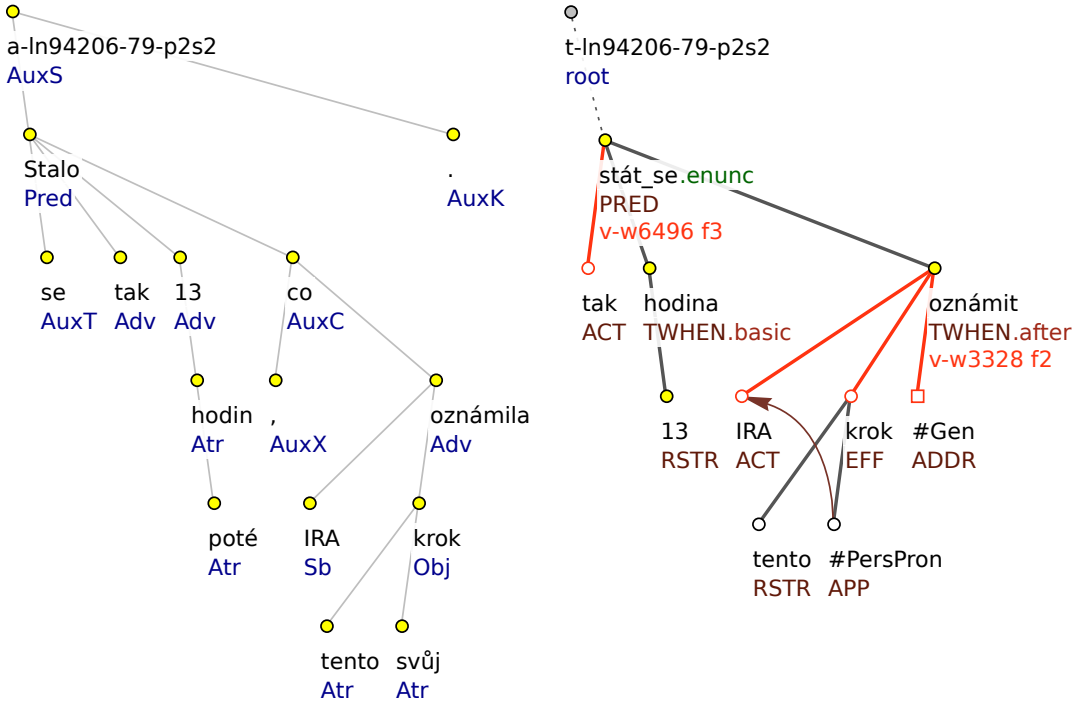
[6] `https://ufal.mff.cuni.cz/pdt2.0`

obrázek 3.2, který už se stal téměř logem PDT. Popis na čtyřech rovinách vypadá takto:



Obrázek 3.2: Stručné schema anotace v PDT. Věta *„Byl by šel do lesa“* včetně překlepu na w-rovině; její oprava, lemmatizace a morfologické značky na m-rovině; její závislostní struktura a analytické funkce na a-rovině; a její hloubková syntax alespoň s údaji o funktorech na t-rovině (včetně doplněného uzlu a nesymetrické korespondence s a-rovinou).

- Nejnižší *w-rovina* dělí vstupní větu na jednotky (slova a interpunkci) – provádí *tokenizaci*. Po technické stránce jim přiřazuje identifikátory, pomocí kterých se s tokeny dále pracuje.
- Na *m-rovině*, též *morfologické* rovině (odpovídající zhruba *morfematické* rovině FGP) jsou opraveny případné chyby vstupního textu a probíhá lemma-



Obrázek 3.3: Skutečná novinová věta a její a-strom (vlevo) a zjednodušený t-strom.
„Stalo se tak 13 hodin poté, co IRA tento svůj krok oznámila.“

tizace a morfologická analýza: každému tokenu je přiřazeno lemma a značka specifikující slovní druh a mnoho dalších morfologických kategorií.

- Následuje *a-rovina* neboli *analytická* rovina (která odpovídá *větněčlenské* rovině FGP). Je to poslední rovina, kde ještě koresponduje pořadí i počet uzlů se slovy ve větě. Rovina zachycuje syntaktické závislostní vztahy mezi větnými členy: podmět, přívlastek, pomocná část slovesa, příslovečné určení apod.
- Nejvyšší *t-rovina* odpovídá v teorii FGP *tektogramatické* rovině neboli rovině jazykového významu.[7] Svou strukturou zachycuje zejména hloubkovou syntax věty, hrany v t-stromě jsou ohodnoceny hloubkovými rolemi (*funktory*) jako například konatel (ACT), materiál (MAT), výjimka (RESTR) a mnoho dalších. Uzly pak nesou informaci lexikální (`t_lemma`) a gramatickou (`gramatém`). Všechny pomocné uzly z a-roviny jsou vypuštěny a většinou nahrazeny vhodnými atributy u významových uzlů; ty korespondují se

[7] V této práci budeme dvojice pojmů t-rovina a tektogramatická rovina směšovat, neboť aspekty, v nichž se liší, nejsou pro tuto práci zásadní.

svými protějšky na a-rovině. Dále jsou doplněny uzly, které v povrchovém vyjádření věty nebyly vyjádřeny, ale do věty nutně patří.
Kromě zmíněného zachycuje t-rovina např. gramatickou koreferenci, slovesnou kontrolu (viz stranu 28), aktuální členění, valenci (přiřazuje lexikální jednotku na základě valenčních doplnění), dále některé jednotlivosti (jména osob, analytické predikáty, slovesné fráze, cizí slova, …).

Z technického hlediska jsou jednotlivé roviny uloženy odděleně, jedna věta se nachází postupně ve čtyřech souborech, které jsou navzájem prolinkované. Co soubor, to jeden celý dokument (několik vět) na jediné rovině.

Napsali jsme, že PDT 2.0 obsahuje 2 000 000 slov. Všechna jsou anotována na w-rovině a m-rovině, tři čtvrtiny dat mají i a-rovinu a 800 000 slov je anotováno na všech rovinách včetně nejhlubší t-roviny.

Na závěr ukážeme skutečnou větu z PDT. Protože jedním z hlavních témat této práce je valence (definice a vysvětlení v sekci 4.1), i v tomto příkladu se na valenci zaměříme. Na obrázku 3.3 vidíme krátkou větu: vlevo je její reprezentace na a-rovině, vpravo její reprezentace na t-rovině. Jde sice skutečně o reálnou větu, nicméně z důvodu čitelnosti stále nezobrazujeme ani zdaleka všechny informace, které se k jednotlivým uzlům vztahují.

Ve stromě na t-rovině je červeně vyznačena valence obou sloves. Sloveso *stát se* zde používá třetí slovesný rámec (f3), který vyžaduje jako povinné jen jediné doplnění, a to ACT v nominativu, nebo ve formě *tak*. Druhé sloveso *oznámit* je zajímavější, neboť rámec f2 vyžaduje tři doplnění: ACT (*kdo*), EFF (*co*), ADDR (*komu* oznamuje). Ovšem ADDR ve větě chybí – IRA svůj krok neoznámila nikomu konkrétnímu. Jedná se tedy o tzv. všeobecného adresáta a dochází zde k doplnění nového uzlu, jak jsme zmínili výše, což je vyznačeno jeho hranatým tvarem. Získává pomocné `t_lemma` #Gen. Do valenčního rámce tohoto slovesa sice ještě patří PAT (oznámit *o něčem*), není ovšem povinný, proto jeho absenci není potřeba nijak řešit.

Ve a-stromě vpravo vidíme podobnou strukturu, kde ovšem vlastní uzel mají i částice *se*, příslovce *poté*, spojka *co* nebo interpunkce. A samozřejmě ohodnocení uzlů je povrchově syntaktické (například *krok* je `Obj`, zatímco na t-rovině jde o EFF).

3.4 PDT 2.5, PDT 3.0

V této práci budeme hovořit též o novějších verzích PDT, konkrétně 2.5 (Bejček et al., 2012) a 3.0 (Mikulová et al., 2013). Skládají se z týchž textů (tedy shodný rozsah korpusu), ale jsou rozšířené o další anotace, převážně na t-rovině. Proto

jde-li nám o samotné věty – mluvíme tedy o datech PDT– je jedno, kterou verzi zmíníme; často dokonce nebudeme v takovém případě uvádět verzi žádnou.

PDT 2.5[8] přidává zejména hranice klauzí na a-rovině a vyznačení víceslovných výrazů (VV) na t-rovině (o nichž mluvíme v sekci 6.1.3). Anotace víceslovných výrazů na t-rovině umožňuje přiblížit se záměru FGP a celý VV (do verze 2.0 reprezentovaný více uzly) sloučit do uzlu jediného.[9] O víceslovných výrazech tedy hovoříme buď jako o součásti PDT 2.5, nebo jako o projektu Lexemann (nad PDT 2.0) – obojí se stejným významem.

PDT 3.0[10] je poslední verze PDT, která obsahuje vše z verzí předchozích a navíc doplňuje širokou škálu koreferenčních a diskurzních vztahů, anotaci žánru dokumentu a další.

3.5 PCEDT

Prague Czech-English Dependency Treebank[11] (PCEDT 2.0, Hajič et al., 2012) je paralelní, ručně anotovaný syntaktický korpus. Rozsahem 1,2 milionu slov (pro jeden jazyk) je sice trochu menší než PDT, ovšem všechna slova mají svou t-rovinu (jakkoli je v PCEDT trochu zjednodušená a ochuzená), takže hloubková anotace pokrývá větší množství slov než PDT.

Data pocházejí ze sekce WSJ v anglickém Penn Treebanku. Složkové stromy neobsahují žádnou vnitřní strukturu jmenných frází (NP), ta byla doplněna s využitím práce Vadas a Curran (2007). Pak mohla být anglická data automaticky převedena ze složkových do závislostních stromů a ručně doplněna o t-rovinu. Česká část vznikla překladem WSJ a následně automatickou metodou vznikla m-rovina a a-rovina, ručně potom t-rovina. Na každé rovině jsou obě jazykové verze spolu provázány odkazy na úrovni jednotlivých uzlů (včetně odkazů z anglické a-roviny i t-roviny na původní složkový strom).

[8] `https://ufal.mff.cuni.cz/pdt2.5`

[9] Přesněji řečeno, PDT 2.0 a další umožňují tři mody zobrazení VV v editoru TrEd: bez znázornění VV, ohraničená oblast kolem celého podstromu VV a konečně zmíněné kompaktní zobrazení, kdy je celý podstrom reprezentován jediným uzlem pro VV.

[10] `https://ufal.mff.cuni.cz/pdt3.0`

[11] `https://ufal.mff.cuni.cz/pcedt2.0/`

4

Slovníky a jejich formáty

Stěžejním tématem této disertace je práce se slovníky a slovníkovými databázemi. Ze širokého spektra slovníků[1] jsme se zaměřili zejména na úzký výsek slovníků současné češtiny, které jsou přístupné v elektronické podobě a v nichž jsou zachyceny (v některých více a v jiných méně) zejména tyto informace:

- **výčet významů lexému** (například výkladové slovníky, ale nejen ty, delimitují homonymní slova do jednotlivých lexikálních jednotek, například $pole_1$ ~ zemědělská půda, $pole_2$ ~ území (minové, naftové, bitevní p.), $pole_3$ ~ hřiště, část herního plánu, $pole_4$ ~ kde působí síla (gravitační, magnetické p.), $pole_5$ ~ okruh činnosti),[2]
- **víceslovné výrazy** (slovníky, jejichž slovníková hesla jsou tvořena více než jedním slovem) a
- **valence slov** (takzvané slovníky valenční, které pro slovníkové heslo, primárně sloveso, popisují, s jakými dalšími částmi věty se pojí, jaké k nim má vztahy a jaká další omezení se na ně vztahují).

Slovníkem (lépe lexikální databází), který si klade za cíl zachytit významy slov, je **WordNet** (část 4.5), v němž jsou slova podle svých významů sdružena do tzv. *synsetů*. Každý synset zhruba vyjadřuje jeden význam, slova v synsetu jsou tedy vzájemně synonymní. Homonymní slova se pak vyskytují ve více různých synsetech. WordNet dále obsahuje řadu sémantických vztahů mezi synsety, z nichž nejvýznamnější je relace hyperonymie. Slovesné synsety obsahují též jednoduché slovesné rámce. WordNet obsahuje též některé víceslovné výrazy (například v české verzi Pala et al., 2010 najdeme výraz *úmorná práce* v synsetu 00366144-n spolu se slovy *otročina*, *nádeničina*, *dřina*, *galeje*, *lopota*, *šichta* a *fuš-*

[1] Například vícejazyčný překladový slovník, jednojazyčný výkladový, pravopisný; tištěný, elektronický, webový; současný, historický, dialektologický, etymologický; retrográdní, frekvenční, technický, biblický, slovník zkratek, synonym, rýmů, …

[2] Podle SSČ (2005), zjednodušeno.

ka, nebo pravděpodobně chybné slovní spojení *kytovec ozubený* samotné v synsetu 01405506-n), ale jejich popis není cílem WordNetu.[3]

SemLex (Straňák, 2010) obsahuje naopak výhradně víceslovné výrazy (VV) (některé dokonce převzaté z českého WordNetu) a jejich význam popisuje tradičním prostředkem: glosou. Víceslovné výrazy však nebývají homonymní, a tak neobsahuje[4] stejný výraz se dvěma významy. Některá synonyma jsou v SemLexu značena odkazem z jednoho slovníkového hesla na heslo synonymní. O dalším propojování hesel viz sekce 7.2.2.

Rozkročeny přes celé toto spektrum stojí valenční slovníky **VALLEX**, **PDT-Vallex** i lexikální databáze **FrameNet**, které pochopitelně zachycují valenci, dá se na ně také velice dobře nahlížet jako na slovníky jednotlivých významů slov (slovníkové heslo je na základě svých zejména syntaktických vlastností delimitováno do jednotlivých významů) a velice detailně popisují i několik typů víceslovných výrazů – v případě zmíněných slovesných slovníků zejména ustálené slovesné fráze, slovesné idiomy a tzv. analytické predikáty. Často se mezi VV počítají také předložkové vazby, frázová slovesa apod. (viz mimo mnohé další např. Sag, Baldwin, Bond, Copestake a Flickinger, 2002), které jsou detailně zachyceny právě ve valenčních slovnících.

V této kapitole se budeme zabývat slovníky a jejich datovými formáty. Zaměříme se zejména na slovníky se syntaktickou informací, které se blíže týkají této práce (jde zejména o slovníky výše uvedené jako příklady), ale představíme i mnoho dalších, které s nimi souvisí, a poskytují tak ucelenější obrázek pestré situace v této oblasti. Mnohé představené slovníky pracují s pojmem *valence*, jí tedy bude věnována první sekce 4.1. Poté přejdeme k popisu jednotlivých slovníků a postupně představíme VALLEX (4.2), PDT-Vallex (4.3), další české valenční slovníky Slovesa pro praxi, Slovník slovesných, substantivních a adjektivních vazeb a spojení, BRIEF, VerbaLex a Český syntaktický lexikon (4.4). Potom přejdeme k lexikálním databázím a popíšeme WordNet (4.5), FrameNet (4.6), a další anglické slovníky PropBank, NomBank a VerbNet (4.7) a jejich propojení v pro-

[3] Ze zhruba 23 000 synsetech českého WordNetu necelých 6 000 obsahuje víceslovný výraz a tyto výrazy tvoří 20 % všech položek slovníku.

[4] Při anotaci VV v PDT 2.0 jsme narazili jen na dva případy víceznačných víceslovných výrazů. Ani jeden se však nakonec do slovníku SemLex nedostal v obou významech, neboť výsledný slovník obsahuje pouze výrazy vyskytující se v PDT. Šlo o sousloví *přímá volba*: jednou ve významu politického procesu volení zástupců, podruhé jako zkratka na mobilním telefonu, kdy jediné tlačítko rychle vyvolá předem nastavenou volbu z menu. Druhé takové sousloví byla *nová vlna*, což může to být umělecké hnutí, zejména v kinematografii, ale také se tak nazývá čistá střižní ovčí vlna.

jektu SemLink (4.8). Posledním představeným slovníkem bude bude opět český slovník víceslovných výrazů SemLex (4.9).

S výjimkou prvních dvou slovníků obsahuje každá sekce část věnovanou formátu slovníku. Formáty VALLEXu a PDT-Vallexu představíme detailně v dalších sekcích (4.10 a 4.11), neboť jejich provázání tvoří stěžejní část této práce. Vysvětlíme výhody i nevýhody zvolených formátů, načež je na ukázkách obou formátů porovnáme v sekci 4.12 a navrhneme nový společný formát SAMR (4.13), který řeší některé dílčí problémy a zejména sjednocuje zápis společných údajů. Jeho vytvoření obhájíme srovnáním s univerzálními formáty (4.14).

4.1 Valence

Nabízíme zde základní vysvětlení pojetí valence ve *Funkčním generativním popisu* (FGP, Sgall, 1967; Sgall et al., 1986, valenční teorie viz především Panevová, 1974, 1980, 1994), jak je uplatněno ve slovnících VALLEX (následující sekce 4.2) i PDT-Vallex (4.3). Chceme zde však podat obecnější, na konkrétním slovníku ještě nezávislý popis, který čerpáme zejména z Lopatková (2010) a Lopatková, Žabokrtský a Kettnerová (2008), kde je také možno hledat další podrobnosti. Z výkladu vypouštíme některé aspekty nepodstatné pro naši práci.

Pojem *valence* jako první použil Lucien Tesnière (1959) a založil na něm svou závislostní gramatiku. Zahájil tím studium a popis valence trvající dodnes. Valenci chápeme jako schopnost plnovýznamového slova[5] otevřít určitý počet pozic pro syntakticky závislé členy (Lopatková, 2010). V této práci nás bude zajímat valence z pohledu hloubkové syntaxe (tektogramatické roviny v terminologii FGP), takže tyto členy, též *valenční doplnění*, budou popisovat tzv. funktory, např. Actor (ACT), Patiens (PAT), nebo jedno ze způsobových určení (MANN, MEANS, CRIT). (Tato terminologie je použita v obou českých valenčních slovnících, s nimiž pracujeme.) Funktor vystihuje druh vztahu mezi řídícím a závislým členem.

Pět prominentních funktorů je nazýváno *aktanty*; jsou to ACT, PAT, EFF (výsledek děje), ADDR (adresát) a ORIG (původ).[6] Někdy se též nazývají *vnitřní doplnění*, jsou to obvykle účastníci popisovaného děje. Jejich určující vlastností je, že se pro jedno slovo nemohou opakovat (pochopitelně kromě případů koordi-

[5] Nejčastěji se hovoří o valenci sloves. Valence sloves je nejpestřejší (ze všech slovních druhů připouští nejširší škálu doplnění stejně jako jejich morfematických realizací), a proto také její zachycení ve slovníku přináší nejvíce informace. Valence slovesa – centra věty – navíc zakládá větnou strukturu. Zpracovává se i valence substantiv, adjektiv či adverbií, ovšem v této práci se jinou než slovesnou valencí nezabýváme.

[6] Pro rychlou představu o použití jednotlivých funktorů si pomůžeme známou větu Jarmily Panevové *„Maminka*$_{\mathsf{ACT}}$ *přešila dětem*$_{\mathsf{ADDR}}$ *loutku*$_{\mathsf{PAT}}$ *z kašpárka*$_{\mathsf{ORIG}}$ *na čerta*$_{\mathsf{EFF}}$*.“*

nace či apozice) a jejich kombinace je specifická pro konkrétní sloveso. Aktanty jsou obvykle rekční a typicky korespondují se subjektem a objekty sloves.

V opozici k vnitřním doplněním stojí *volná doplnění*, která jsou sémanticky distinktivní a každé z nich se u slova může vyskytnout opakovaně, např. *„Až v neděli*$_{\text{TWHEN}}$ *dopoledne*$_{\text{TWHEN}}$ *3. září*$_{\text{TWHEN}}$ *v 11 hodin*$_{\text{TWHEN}}$ *oznámila britská vláda světu, že od této chvíle se považuje ve válečném stavu s Německem.“* — [PDT] Volná doplnění obvykle upřesňují průvodní okolnosti popisovaného děje.[7]

Třetí, novější kategorie *kvazivalenčních* doplnění (Lopatková a Panevová, 2006) stojí na hranici mezi dvěma právě představenými. Stejně jako aktanty není možné opakovat ani kvazivalenční doplnění a také ona jsou charakteristická pro určité sémantické skupiny sloves. Jsou ovšem sémanticky distinktivní a obvykle jsou nepovinná.

Tím se dostáváme k vlastnosti obligatornosti. *Obligatorní* je takové valenční doplnění, jehož přítomnost v tektogramatické struktuře věty je povinná a lze ho elidovat pouze z povrchové realizace věty (kritériem je tzv. *dialogový test*, Panevová, 1974). Protikladem je *fakultativní* doplnění, které nemusí být ve významové rovině věty přítomno; to je obvykle případ většiny volných doplnění.

Množina valenčních doplnění charakteristických pro jeden z významů slova se nazývá *valenční rámec*. Korespondence mezi významy slova a odlišnými valenčními rámci je složitější, dá se ale říci, že typicky skutečně vymezení dvou různých valenčních rámců (tedy hloubkově syntaktických vlastností) je ve shodě s delimitací dvou odlišných významů.

V rámci FGP je zvykem množinu valenčních doplnění řadit podle tzv. *systémového uspořádání*: Sgall et al. (1986) ho definují jako pořadí doplnění v samostatně stojící bezpříznakové větě.

Závěrem ještě musíme zmínit *princip posouvání* (Panevová, 1980), který se aplikuje na aktanty. Je-li ve valenčním rámci slovesa přítomen pouze jeden aktant, je to z definice vždy ACT. Případný druhý aktant je vždy PAT. Zcela se tak odhlíží od sémantického rysu aktantů. Teprve má-li sloveso tři a více aktantů, nastupuje rozhodnutí podle smyslu doplnění – přesto však stále tak, že ACT a PAT nemohou být vynechány.

[7] Z 24 funktorů pro volná doplnění použitých ve VALLEXu i PDT-Vallexu zmiňme pro představu např. příčinu (CAUS), určení směrová (DIR1, DIR2, DIR3) a místní (LOC), způsob (MANN) a prostředek (MEANS) nebo různé typy časových určení (TFHL, TFRWH, THL, TOWH, TSIN, TTIL, TWHEN).

4.2 VALLEX

Slovník VALLEX ve své dnešní podobě začal vznikat v Ústavu formální a aplikované lingvistiky počátkem nového tisíciletí, ovšem na základě mnohem starší teorie FGP, jejíž valenční část vychází zejména z Tesnièra, ale reflektuje také přístupy Fillmorovy (Fillmore, 1968, 1977). Obecné vlastnosti valence i její konkrétní pojetí ve FGP (obojí popsané v předešlé sekci) se promítly do slovníku VALLEX.

V této práci budeme používat slovník VALLEX ve verzi 2.5, který vyšel v roce 2007 jako data (Lopatková et al., 2007) a ve verzi 2.6, který je rozšířen pouze o propojení s PDT-Vallexem (sekce 5.1). Čtenář tedy nemusí mezi těmito dvěma verzemi rozlišovat. Změny, kterých slovník od té doby dodnes doznal, se týkají oprav a rozšiřujících informací – verze 2.5 však již měla stejný rozsah, tedy počet slovníkových hesel (=lexémů) jako dnes.

VALLEX tedy popisuje vlastnosti 2 730 výhradně slovesných lexémů, které jsou rozděleny na 6 460 lexikální jednotek (LU z anglického „lexical unit“). Ty zhruba odpovídají jednotlivým významům sloves. Slovesa byla pro VALLEX vybírána na základě jejich četnosti v SYN 2000[8] a prvních zhruba 2 500 slovesných lemmat[9] bylo doplněno na kompletní lexémy, což v pojetí FGP obnáší všechny vidové protějšky a ortografické varianty, pokud existují.[10] K těmto lexémům byly přidány reflexivní varianty,[11] čímž se počet lexémů rozšířil na zmíněných 2 730.[12]

[8] *Český národní korpus – SYN2000.* Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: `http://www.korpus.cz`.

[9] Přesněji *m-lemmat*, jak nazýváme slovesnou část případně reflexivního slovesa. Mezi těmito 2 500 m-lemmaty tak bylo např. sloveso *brát*, ale nikoli už *brát si*, či brát se.

[10] Pokud tedy mezi prvními 2 500 lemmaty bylo sloveso *doprovázet*^{impf} i sloveso *doprovodit*^{pf}, jsou popisovány v rámci jednoho lexému. Pokud se do slovníku také dostalo sloveso *donutit*^{pf}, ale kritériem četnosti již neprošel jeho protějšek *donucovat*^{impf}, byl do slovníku přesto zařazen i tento protějšek, aby lexémy byly úplné. Podobně byla zařazena celá řada *iterativ*, například lexém *doléhat*^{impf}, *dolehnout*^{pf}, *doléhávat*^{iter}.

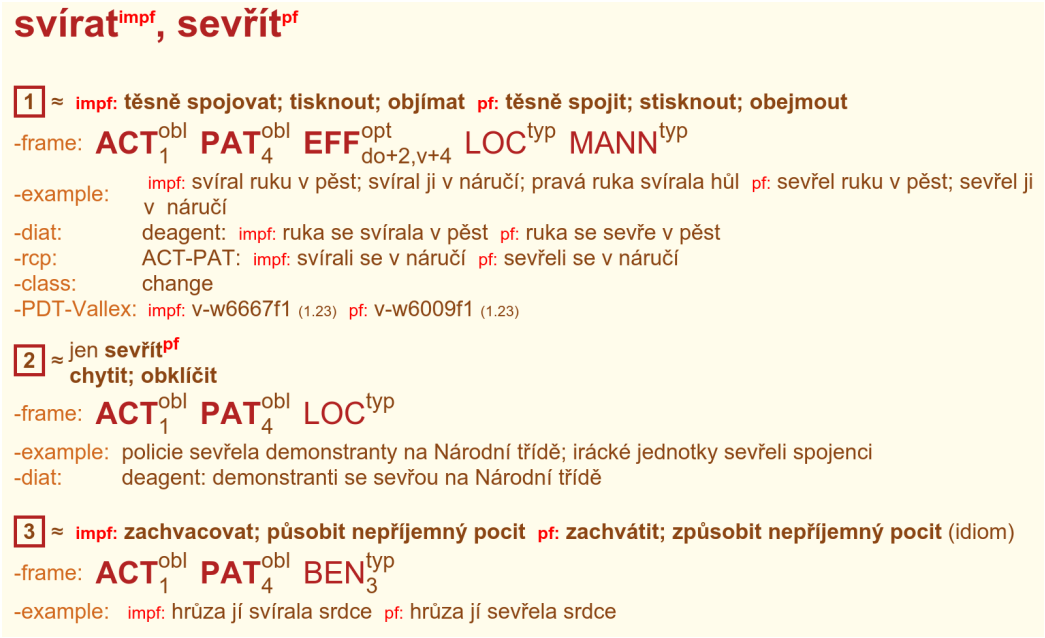
[11] Pouze primární reflexiva tantum (*smát se, chovat se*) a odvozená (sekundární) reflexiva (*vrátit se*) tvoří samostatné lexémy. Pouhé reflexivní užití nereflexivních lexémů je zachyceno u nereflexivního lexému, například jako informace o možnosti tvořit reflexivní pasivum, či reciproční konstrukci.

[12] Kromě přidaných reflexiv byla ještě některá lemmata rozdělena na dva různé *homografy*, pokud se lišily výrazně například etymologicky jako *dopouštět-I* ~ způsobovat a *dopouštět-II* ~ naplňovat vodou. Potom tedy vytvořila dva lexémy. V tomto případě společně se svými vidovými protějšky, které jsou také homonymní: po řadě *dopustit-I* a *II* se stejným významem. Ale nemusí tomu tak být: *olupovat-I* tvoří lexém s *oloupit* a *olupovat-II* lexém s *oloupat*.

Podrobnější informace o vzniku slovníku i o jeho struktuře jsou v úvodu Lopatková et al. (2008). V sekci 5.1 naší práce jsou pak uvedena některá přesná kvantitativní data o VALLEXu (ve srovnání s PDT-Vallexem) v tabulkách 5.1 a 5.2.

Při sestavování jednotlivých hesel VALLEXu, zejména při delimitaci lexémů do jednotlivých lexikálních jednotek, byly využívány tyto slovníky a korpusy: BRIEF (Pala a Ševeček, 1997), SSČ, SSJČ, Slovesa pro praxi (Svozilová, Prouzová a Jirsová, 1997), Slovník slovesných, substantivních a adjektivních vazeb a spojení, ČNK a PDT 2.0 (reference viz Lopatková et al., 2008, str. 9).

Z technického pohledu a značně do hloubky rozebereme VALLEX v následujících sekcích, kde se budeme zabývat datovým formátem slovníků.[13] Nyní se však přesto sluší čtenáře se základními vlastnostmi slovníku seznámit. Představíme je na příkladu dvou sloves, *„sevřít, svírat“*, na obrázku 4.1.



Obrázek 4.1: Ukázka slovníkového hesla *„sevřít, svírat“* z webového rozhraní slovníku VALLEX.

[13] Vyčerpávající popis informací obsažených ve webové verzi slovníku je na adrese `http://ufal.mff.cuni.cz/vallex/2.5/doc/structure_cs.html` dostupné odkazem ze slovníku.

Každé slovníkové heslo odpovídá celému **lexému**[14] (více při popisu formátu v bodu 2 na straně 61), v našem případě sdružuje dvě lemmata: nedokonavé (imperfektivní) *svírat* a dokonavé (perfektivní) *sevřít*. Následují tři očíslované lexikální jednotky společné pro všechna lemmata ze záhlaví, není-li uvedeno jinak (což je případ druhé LU, která se nevztahuje na sloveso *svírat*).

Každá lexikální jednotka obsahuje **glosu** (typicky zvlášť pro každý vidový protějšek), která stručně vystihuje přibližný význam slovesa, je-li použito s uvedenými valenčními doplněními, což usnadňuje práci se slovníkem.

Krom obvyklých významů sloves, které by měly být pokryty plně, jsou do slovníku zařazeny také některé idiomatické lexikální jednotky; takové jsou v záhlaví opatřeny značkou „idiom“.

Následuje nejdůležitější součást LU, valenční rámec, který je pro příslušná lemmata *vždy* společný bez obměn. Valenční rámec sestává z jednotlivých valenčních doplnění, jak byly popsány v předchozí sekci 4.1. Uváděny jsou obligatorní i fakultativní aktanty stejně jako kvazivalenční doplnění. Po nich následují obligatorní volná doplnění. Nad rámec teorie FGP jsou na konci valenčního rámce uváděny ještě další volná doplnění, která sice nejsou obligatorní, ale jsou tzv. typická pro danou LU: vystihují její význam, nebo se u ní objevují často (obvykle jsou charakteristická pro celou sémantickou skupinu sloves). Jejich pořadí je motivováno systémovým uspořádáním. Aktanty jsou řazeny takto: ACT ADDR PAT ORIG EFF. Jsou uváděny na začátku rámce, za nimi následují všechna další obligatorní doplnění, potom fakultativní kvazivalenční a na závěr typická volná doplnění. Pořadí v rámci těchto skupin je pevně dané. Pro obligatorní rámce se používá zkratka „obl“, pro fakultativní „opt“ a pro typické „typ“.

Zkratka pro funktor může být v dolním indexu následována explicitně zachyceným morfematickým vyjádřením. Je to výčet některých z těchto typů hodnot:

- bezpředložkové pády (např. 1 = nominativ; 7 = instrumentál)
- předložkové skupiny (např. z+2 = předložka *z* následovaná genitivem)
- infinitivní konstrukce (inf)
- závislé věty (např. že = vedlejší věta uvozená spojkou *že*; cont = obsahová vedlejší věta uvozená např. tázacím zájmenem)
- konstrukce s adjektivy (např. adj-4 = adjektivum v akuzativu)
- konstrukce s *být* (např. být+adj-7 = *být* následované adjektivem v instrumentálu; být+7 = *být* následované substantivem v instrumentálu)
- část frazému (např. napospas; čistého vína)

[14] Případné varianty jsou odděleny lomítky. Pokud se ve slovníku vyskytují stejná lemmata v různých lexémech, výše zmíněné homografy, jsou odlišeny římskou číslicí.

Pro aktanty je takový výčet úplný, jiné formy nejsou přípustné. Pokud morfematické formy nejsou uvedeny, může se použít libovolné (implicitní) doplnění typické pro daný funktor.

Příležitostně budeme v této práci pro oba valenční slovníky používat následující přepis (zde ukázka pro první lexikální jednotku), který využívá závorky pro zápis přípustných morfematických forem: ACT(1) PAT(4) EFF(do+2,v+4)^{opt} LOC^{typ} MANN^{typ}. Obligatornost explicitně neuvádíme.

Po glose a valenčním rámci obsahuje každá lexikální jednotka (introspektivní) příklad. Od verze 2.6 je pro vybraná slovesa k dispozici také řada příkladů z ČNK (díky projektu VALEVAL a propojení se starší verzí slovníku, viz sekce 4.2.1 a 5.2).

Tím jsme vyjmenovali povinné součásti každé lexikální jednotky. Dále mohou následovat rozšiřující syntaktické informace, jsou-li pro danou LU přípustné. Jsou to deagentní diateze, reflexivita,[15] reciprocita, kontrola a syntakticko-sémantická třída.

Deagentní diateze (značená jako „diat“) zachycuje možnost LU vstupovat do tzv. sekundární diateze: příznakové konstrukce tvoří tranzitivní slovesa (např. *„Starší lidé se často odstrkují ve prospěch mladších.“*, hodnota „deagent“), i intranzitivní (např. *„Ke křenu se nevoní!“*, hodnota „deagent0“). Diateze je vždy doplněna ilustračním příkladem. V ukázce na obrázku 4.1 ji umožňují LU 1 a 2.

Atribut reflexivity (značený jako „rfl“) je přidělen těm LU, které umožňují jeden z aktantů vyjádřit zájmenem *se/si* koreferenčním se subjektem, tedy například *„Vyhradil si právo na užívání příjezdové cesty.“* (ADDR vyjádřený reflexivním zájmenem *si* je v dativu, hodnota atributu je „cor3“), *„Trávila se z nešťastné lásky.“* (PAT vyjádřený zájmenem *se* je v akuzativu, hodnota atributu je „cor4“). Informace o typech reflexivity je opět opatřena příklady.

Reciprocitou (viz Panevová, 2007; Panevová a Mikulová, 2007) se rozumí možnost symetrického použití dvou (a více) valenčních doplnění dané lexikální jednotky, například ACT(1) OBST(o+4)^{opt} se místo *„Zuzana*$_{ACT}$ *se o Honzu*$_{OBST}$ *opřela.“* změní na *„V neštěstí se o sebe opřeli.“* Ne každá lexikální jednotka toto umožňuje (a navíc se reciprocita může rozšířit i na tři funktory: ACT+ADDR+PAT *„Namluvili si o sobě lži.“*.) Ve VALLEXu je reciprocita reprezentována atributem „rcp“ následovaným výčtem funktorů, které lze nahradit reciproční konstrukcí, a příkladem. Pro první LU na obrázku 4.1 jsou to ACT s PAT.

Slovesa *kontroly* mají jedno valenční doplnění reprezentované infinitivem (viz Panevová, 1996). Podmět tohoto infinitivu, který je povrchově nevyjádřený, je

[15] Ve verzi 2.5 byly ještě deagentní diateze i reflexivita slučovány pod jeden společný atribut. Později však došlo k přejmenování (bez jiných změn) a zde se tedy přidržíme aktuální terminologie.

člen kontrolovaný (*controlee*). Kontrolující člen (*controller*) je výraz typicky vyplňující jednu z valenčních pozic slovesa kontroly, který je s členem kontrolovaným koreferenční. Mezi kontrolujícím členem a kontrolovaným členem existuje vztah kontroly. Atribut „control“ má potom hodnotu funktoru kontrolujícího členu; v případě, že takový člen valenčního rámce neexistuje, ale reciprocita je umožněna, má atribut hodnotu „ex“.

↪ *Příklad kontroly:* V krátké větě *„Aneta se bojí létat.“* je controller sloveso *bát se*, jeho ACT je *Aneta*. Nevyjádřený a nevyjádřitelný podmět kontrolovaného slovesa *létat* je taktéž *Aneta*, tedy ACT slovesa kontrolujícího. V tomto případě by tedy bylo „control: ACT“.

Téměř polovině lexikálních jednotek byla přiřazena syntakticko-sémantická třída. Jde o pracovní klasifikaci LU. Různé jednotky v rámci jednoho lexému mohou spadnout do různých tříd. Hodnoty atributu „class“ jsou například „change“ (pro *proměňovat, klesat, růst, …*, či naše ukázkové $svírat_1^{impf}$, $sevřít_1^{pf}$), „motion“ (pro *běžet, dorážet, hýbat se, …*), „psych verb“ (pro *klamat, potěšit, polekat, …*), nebo „transport“ (pro *přemisťovat, donášet, shrnovat, …*).

Od verze 2.6.2 (leden 2014) je lexémům a také vybraným lexikálním jednotkám přiřazen odkaz do slovníku PDT-Vallex (viz sekce 5.1), jak je pro první LU vidět i na obrázku 4.1.

XML formátu VALLEXu se budeme věnovat podrobně v samostatné sekci 4.10 po představení ostatních slovníků.

4.2.1 VALEVAL

Na slovníku VALLEX je založen projekt VALEVAL (Bojar, Semecký a Benešová, 2005), který zde představíme též, neboť s ním budeme pracovat v kapitole 5.2 a zejména 6.3.

VALEVAL poskytuje anotovaná data k vybraným slovesům z VALLEXu ve verzi 1.0. Každé z těchto sloves je v textu z ČNK SYN 2000 označeno jednou z LU.[16]

Pro manuální anotaci bylo ze slovníku vybráno 109 m-lemmat tak, aby rovnoměrně zastupovala „obtížná“ i „snadná“ slovesa (měřeno počtem LU). Byly

[16] V rámci projektu VALEVAL se také testovalo automatické přiřazování lexikálních jednotek k výskytům sloves v textu. Jak jsme už řekli, LU odpovídají významům, jde tedy vlastně o desambiguační úlohu (Word Sense Disambiguation). Odtud také název: cílem VALEVALu byla desambiguace valenčních rámců (LU) podobně jako cílem SENSEVALu byla desambiguace významů.

vybírány celé shluky se stejným m-lemmatem[17] (tj. m-lemma společně s případnými reflexivními slovesy i s vidovými protějšky).

Ke každému m-lemmatu bylo automaticky vybráno 100 vět z ČNK. Těchto sto vět pro každé z m-lemmat (spolu s kontextem předchozích tří vět) dostali tři anotátoři. Jejich úkolem bylo přiřadit:

- sloveso ze slovníku (tedy celé slovníkové heslo, neboť ve verzi 1.0 ještě nebyly sdruženy vidové protějšky, ani reflexivní varianty),
- lexikální jednotku (valenční rámec), přičemž mohli
  - v případě chybně vybrané věty značit, že se jedná o jiné sloveso (0,6%),[18]
  - v případě nesrozumitelného kontextu tuto skutečnost označit (0,3%),
  - v případě nejistoty přiřadit více LU (2,3%) a
  - v případě, že potřebná LU ve VALLEXu chyběla, označit její absenci (5,4%),

  a
- navíc mohli své anotaci přiřknout značku nejistoty (3,4%).

Každá věta byla zpracována paralelně všemi třemi anotátory. Jejich anotátorská shoda je v tabulce 4.1 Jedno z překvapivých zjištění zmíněného článku Bojar et al. (2005) je, že shoda anotátorů nesouvisí ani s frekvencí slovesa (v ČNK) ani s jeho komplexností (měřeno opět počtem LU).

	IAA [%]		κappa	
	ø	vážený ø	ø	vážený ø
Všechny věty	66,8	61,4	0,52	0,54
Věty bez nejistoty	68,2	73,7	0,58	0,62

Tabulka 4.1: Mezianotátorská shoda a kappa tří paralelních anotací v projektu VALEVAL pro 109 sloves. Druhý řádek je pouze pro takové věty, kde žádný z anotátorů nezvolil příznak nejisté anotace. Vážené hodnoty jsou vážené počtem výskytů daných sloves v ČNK.

Pro tuto práci je nejdůležitější, že výstupem VALLEXu byla „gold data“: 8 066 vět pro 107 m-lemmat, kde se všichni tři anotátoři přesně shodli (případně v nichž byly dodatečně opraveny drobné chyby při čištění dat). Těchto sto sedm m-lemmat nemá samozřejmě přiřazenu větu pro každou ze svých LU, neboť data byla vybírána náhodně, tedy jsou reprezentativní a nepokrývají nutně všechny

[17] Ve VALLEXu 2.0 později nazývané lexémové shluky.

[18] Čísla vyjadřují, v kolika procentech případů byla daná anotační možnost využita.

významy slovesa. Přesto to však jsou významná (jediná a nikoli malá) anotovaná data pro VALLEX. Nevýhodou ovšem je, že byla navázána na VALLEX 1.0, jehož lexikální jednotky prošly do verze 2.0 významnými změnami. Vrátíme se ke „golden VALEVALu“ v sekci 5.2, kde popíšeme, jak jsme data zpřístupnili propojením lexikálních jednotek mezi VALLEXem 1.0 a 2.0.

4.3 PDT-Vallex

Druhý český valenční slovník, s nímž budeme pracovat, se nazývá PDT-Vallex (Hajič et al., 2003; Urešová, 2011a,b). Také vznikal v Ústavu formální a aplikované lingvistiky, a to od roku 2002, proto má podobný název jako VALLEX. Je však úzce propojený s PDT 2.0, s nímž vznikal zároveň a na jehož data jsou hesla PDT-Vallexu navázána. Aby v tomto textu nedocházelo k záměně VALLEXu za PDT-Vallex či naopak, můžeme udělat pouze tolik, že slíbíme, že budeme zásadně psát VALLEX verzálkami a PDT-Vallex takto minuskami, aby se jejich názvy alespoň trochu vizuálně odlišovaly.

Také PDT-Vallex stojí na Funkčním generativním popisu jazyka. Jeho vznik pomohl také ověřovat valenční teorie FGP na datech. Verze 1.0 vznikala společně s PDT 2.0 z důvodu zachování konzistence anotace: vyskytovalo-li se stejné slovo ve stejném významu v PDT vícekrát, bylo nutné zajistit, že jeho valence bude anotovaná vždy stejně. Valenční rámec byl vytvořen zpravidla až při nalezení takového výskytu v anotovaných datech. Také obecný návod na tvorbu hesel z této anotace teprve vyplynul, nikoli aby jí předcházel, což je významný aspekt, který PDT-Vallex odlišuje od většiny slovníků a který hlavní autorka, Urešová (2011a), považuje za největší přednost PDT-Vallexu. PDT-Vallex 2.0 byl potom výrazně rozšířený během anotace PCEDT (Pražského česko-anglického závislostního korpusu). Tato rozšířená verze obsahuje 11 656 hesel (z toho 7 103 slovesných) a pokrývá 17 720 rámců (z toho 11 933 slovesných).

PDT-Vallex zachycuje valenci všech sloves, která se vyskytují v PDT. Dále obsahuje také vybraná substantiva, adjektiva a adverbia a jejich valenční charakteristiku. Od výskytů dotyčných slov v PDT (konkrétně v jeho tektogramatické reprezentaci) vede vždy odkaz na konkrétní valenční rámec ve slovníku. Nás budou v této práci ovšem zajímat výhradně slovesa.

Slovník je zveřejněn na webu[19] a jeho ukázku opět pro slovesa *svírat* a *sevřít* vidíme na obrázku 4.2.

Základní slovníkové heslo je zde – podobně jako ve většině tradičních slovníků – reprezentováno samotným lemmatem[20] (včetně případné reflexivní částice).

[19] `http://ufal.mff.cuni.cz/lindat/PDT-Vallex.html`

[20] Přesněji tektogramatickým `t_lemma`tem.

Obrázek 4.2: Ukázka slovníkových hesel *sevřít* a *svírat* z webového rozhraní slovníku PDT-Vallexvčetně zobrazení jednoho z korpusových dokladů z PDT.

Vidové protějšky tedy tvoří samostatná hesla (která jsme do ukázky vybrali obě dvě).

Slovníkové heslo je vždy rozděleno na jednotlivé číslované rámce (což by ve VALLEXu odpovídalo lexikálním jednotkám), které se typicky liší významem i valenčním rámcem. Rámce mohou mít význam konkrétní, abstraktní, nebo frazeologický. PDT-Vallex zachycuje jen ty informace o rámcích, které byly při anotaci potřebné pro správnou (konzistentní) konstrukci tektogramatického závislostního stromu.

Zápis valenčního rámce je po představení VALLEXu zřejmě srozumitelný: taktéž používá funktory tektogramatické roviny k popisu členů valenčního rámce a u nich uvádí v závorce morfo-syntaktické informace (z analytické, nikoli tektogramatické roviny PDT). Morfo-syntaktické informace se týkají „základní“ formy, tedy aktivního užití slovesa.

V PDT-Vallexu je ovšem možné zachytit mnohem složitější informace o valenčních doplněních, než je vidět na obrázku 4.2, včetně jejich závislostních vzta-

hů. Například frazém *„lámat si hlavu“* je reprezentován takto: ACT(1) DPHR(hlava:4) PAT(7;s+7;↓c;nad+7), což nám říká, že hlava musí být v akuzativu a následuje buď slovo v instrumentálu, nebo slovo v instrumentálu uvozené předložkou *s*, nebo obsahová závislá věta (uvozená vztažným zájmenem či zájmenným příslovcem), nebo slovo v instrumentálu uvozené předložkou *nad*. Složitější rámec, který již obsahuje i syntaktickou informaci, obsahuje například frazém *„ležet na bedrech“*: ACT(1) DPHR(na-1[bedra.6];na-1[bedra.6[.u#]]). Hranaté závorky uzavírají syntakticky závislý podstrom, tedy na předložce *na* (číslice 1 značí, že nejde o citoslovce *na*) závisí *bedra* v lokálu – a na nich ještě (druhá varianta) může záviset přivlastňovací zájmeno nebo přídavné jméno, které se bude s *bedry* shodovat v pádě, čísle a rodě. Vyjadřovací síla je ovšem mnohem větší, než jsme ukázali v těchto příkladech; pro podrobnější popis viz Mikulová et al. (2006, sekce 10.5).

Funktory i záznam syntaktické informace závislých uzlů využívali anotátoři pro udržení konzistence označování vztahů mezi slovesem a závislými členy: vždy stejný funktor pro totéž, vždy stejná vnitřní syntaktická struktura téhož závislého výrazu.

Pořadí členů valenčního rámce je opět pevně dané a řízené příslušnými funktory, nicméně odlišné od pořadí ve VALLEXu. Například jmenné části analytických predikátů a doplnění slovesných frází přichází hned za ACTorem a PAT předchází ADDResáta: ACT CPHR DPHR PAT ADDR ORIG EFF.[21]

Členy rámce označené otazníkem jsou *fakultativní*,[22] všechny ostatní (tedy také všechny v našem příkladu) jsou obligatorní. Počet a druh *obligatorních* doplnění usnadnil při anotaci PDT 2.0 vytvoření nových t-uzlů pro všechny nevyjádřené členy valenčního rámce: technické uzly s `t_lemmatem` #Gen pro aktant a #Oblfm pro volné doplnění. Ve shodě s teorií FGP se žádné jiné funktory v rámci neuvádějí, *typická* volná doplnění však bývají vyznačena v příkladech, pokud se v nich vyskytují – opět z důvodu anotátorské konzistence.

Protože se (až na výjimky) nevyskytují rámce, které se neobjevily v datech, obsahuje každý valenční rámec také informaci o frekvenci, a to zvlášť pro PDT (stojaté písmo, 2x u slovesa *sevřít*) a pro PCEDT (kurzivní písmo, *1x* u *sevřít*). Po kliknutí na počet výskytů se objeví seznam všech příkladů z korpusu obsahující

[21] Uvedená kombinace se nemůže nikdy objevit společně, vždy jen její část. Ta však bude seřazena podle tohoto klíče.

[22] PDT-Vallex nezaručuje, že je u všech rámců uveden úplný výčet fakultativních aktantů, neboť jejich stanovení je obtížné, pokud se v datech PDT nevyskytly. Totéž ostatně platí i pro morfematické realizace jednotlivých členů rámce.

také vyznačení valenčních doplnění (včetně případné koreference). Na obrázku 4.2 je takto zobrazen jediný příklad pro druhý rámec slovesa *svírat*.[23]

Za valenčním rámcem následuje obvykle stručné vyjádření významu daného rámce (synonymum, vidový protějšek apod.) a ukázkové příklady, což sloužilo pro snazší orientaci a rozhodování anotátorů PDT.

PDT-Vallex je uložen ve formátu XML, přesněji v jeho variantě PML, viz sekci 4.11 za přehledem ostatních slovníků, kde se formátům valenčních slovníků budeme věnovat podrobně.

4.4 Další české valenční slovníky

Představené slovníky VALLEX a PDT-Vallex nejsou jediné ani první valenční slovníky pro češtinu. Následující přehled vychází z práce Urešové (2011a).

Slovesa pro praxi

První český valenční slovník, Slovesa pro praxi (Svozilová et al., 1997), vznikal na základě lístkového lexikálního archivu Ústavu pro jazyk český AV ČR a popisuje nejčastější česká slovesa (řazeno podle frekvenčního slovníku češtiny z 60. let, Jelínek, Těšitelová a Bečka, 1961). Každému slovesnému významu je zde přiřazen větný vzorec a valenční analýza, která obsahuje údaje o slovním druhu, tvaru, syntaxi a sémantice valenčních doplnění.

Slovník slovesných, substantivních a adjektivních vazeb a spojení

Tento slovník (Svozilová, Prouzová a Jirsová, 2005) navazuje na Slovesa pro praxi a obsahuje téměř 16 000 českých slov různých druhů, jak je patrné z názvu. Každá lexikální jednotka obsahuje charakteristická slovní spojení (např. *„říkat někomu něco o někom, o něčem“*) která vystihují její význam. Zcela ovšem rezignuje na jakékoliv formální zachycení těchto vazeb.

BRIEF

Slovník BRIEF (Pala a Ševeček, 1997) vychází zejména ze Slovníku spisovného jazyka českého (Havránek et al., 1989) a přebírá z něj informaci o možnostech povrchového vyjádření slovesných doplnění a o jejich možných kombinacích. U slovesných doplnění se kromě již známých informací (pád, předložka, vedlejší

[23] Věty z PDT a PCEDT jsou součástí pouze webové verze, v datových souborech slovníku (sekce 4.11) ani v tištěné verzi (Urešová, 2011b) obsažené nejsou. Jsou dosažitelné z PDT, kde každé sloveso obsahuje odkaz do PDT-Vallexu.

věta či infinitiv, idiom) uvádějí také sémantické rysy (a dále negace, jiné omezení, poznámka).

Na rozdíl od obou slovníků vycházejících z FGP neinformuje BRIEF o hloubkové roli jednotlivých doplnění. Další odlišností je vyčleňování rámců, kdy nový rámec je vytvořen pro každou kombinaci morfematického vyjádření slovesných doplnění, která se vyskytla v příkladech v SSJČ. Totéž platí pro rámce, které se liší jen obligatorností doplnění. Zachycuje tudíž spíše rekci a není zde zřetelná korespondence mezi rámci a významy slovesa. Nezachycuje ani doplnění vyjadřující konatele děje, tedy typicky subjekt věty v nominativu.[24]

Pro zápis slovníku BRIEF slouží dva ekvivalentní formáty: (i) velmi úsporný BRIEF vhodný ke strojovému zpracování a (ii) formát Verbose, který je přístupnější pro člověka. Oba zápisy je možné použít pro vkládání dat do slovníku a jsou na sebe vzájemně převoditelné. Následuje příklad slovesa *dopadat* v obou formátech.

BRIEF:

```
dopadat <v>hTc4-hPTc4r{na},hTc4-hPTc7r{s},hPTc4,hTc6r{při},
          hTc2r{u},hPTc4-hTc6r{při},hPTc4-hTc2r{u}
```

Verbose:

```
dopadat
= co & na koho|co
= co & s kým|čím
= kdo|co
= při čem
= u čeho
= koho|co & při čem
= koho|co & u čeho
```

Kompaktní zápis formátu BRIEF vysvětlíme: Co řádek, to jedno sloveso. Začíná infinitivem a za značkou `<v>` následuje výčet možných slovesných rámců oddělovaných čárkami. Členy rámce jsou oddělovány pomlčkou a sestávají z dvojic atribut (malé písmeno) – hodnota (velké písmeno, nebo složené závorky). Sémantický rys (`h`) nabývá v ukázce hodnoty životné (`P`) a neživotné (`T`). Další zde použité atributy jsou pád (`c`) a předložka (`r`). Při porovnání se zápisem Verbo-

[24] Například morfematická vyjádření následujících tří rámců slovesa *dopadnout*, pokud je přepíšeme zhruba do syntaxe, na kterou jsme zvyklí, vypadají v BRIEFu takto:
ARG2(4)$^{\text{osoba/věc}}$
ARG2(4)$^{\text{osoba/věc}}$ ARG3(při+6)$^{\text{věc}}$
ARG2(4)$^{\text{osoba/věc}}$ ARG3(u+2)$^{\text{věc}}$
Ve VALLEXu či PDT-Vallexu by byla všechna tři vyjádření sloučená do jediné lexikální jednotky a vypada by asi takto: ARG1(1)$^{\text{obl}}$ ARG2(4)$^{\text{obl}}$ ARG3(při+6,u+2)$^{\text{opt}}$.

se, který stejnou informaci rozepisuje za použití zájmen a předložek, je i formát BRIEF srozumitelný.

VerbaLex

Lexikální databáze VerbaLex vychází ze slovníků VALLEX, BRIEF a Slovesa pro praxi. Také VerbaLex – podobně jako BRIEF – akcentuje sémantickou rovinu popisu, k čemuž využívá český WordNet a jeho systém „synsetů“ (viz následující sekci 4.5).

Vycházeje ze zdrojů s morfologickou i syntaktickou strukturou, obsahuje i VerbaLex morfo-syntaktické informace. Ty zobrazuje způsobem srozumitelnějším pro nepoučené čtenáře, totiž zájmennými výrazy reprezentujícími přímé i předložkové pády. Například tedy *třídit*$_3$:

AG(kdo1)$^{\text{obl}}$ VERB$^{\text{obl}}$ ABS$^{\text{obl}}$|ENT(co4)$^{\text{obl}}$ ATTR(podle+čeho2, do+čeho2)$^{\text{obl}}$ GROUP(na+co4)$^{\text{opt}}$

VerbaLex je zapisovaný jako XML, které je blízké XML VALLEXu ve formátu B (sekce 4.10), neboť oba formáty vycházejí z formátu VALLEXu 1.0. Proto také nebudeme formát popisovat, připojíme jen zkrácenou ukázku z Horáka (2008) pro synset {*dodat:1, dát:8, vložit:1, vsunout:1, přidat:2, připojit:1*}:

```
<word_entry>
  <headword_lemmata>
    <lemma ord='1' sense='1' aspect='pf' aspectual_counterpart_lemma='dodávat'>dodat</lemma>
    ...
  </headword_lemmata>
  <frame_entry frame_index='1'>
    <frame_lemmata>
      <lemma sense='8' aspect='pf'>dát</lemma>
      ...
    </frame_lemmata>
    <synonym_lemmata>
      <lemma aspect='pf' sense='1'>vložit</lemma>
      ...
    </synonym_lemmata>
    <example>dok: připojili ke smlouvě své podpisy</example>
    <use>prim</use>
    <frame_slots>
      <slot number='1' functor='AG' type='obl' class='person:1'>
        <form type='direct_case' case='kdo1' />
      </slot>
      <slot number='2' type='obl' functor='VERB'/>
      <slot number='3' functor='INFO' type='obl' class='info:1'>
        <form type='direct_case' case='co4' />
      </slot>
      <slot number='4' functor='COM' type='obl' class='written communication:1'>
        <form type='prepos_case' prepos_lemma='k' case='čemu3'/>
      </slot>
    </frame_slots>
  </frame_entry>
  ...
</word_entry>
```

VerbaLex je bohužel v plné verzi jen těžko dostupný; v současné době je jeho webová demoverze na adrese `http://nlp.fi.muni.cz/verbalex/htmlDEMO/`.

Český syntaktický lexikon

Ambicí Českého syntaktického lexikonu (Skoumalová, 2001) je jeho využití při počítačovém zpracování češtiny (desambiguace, parsing, tektogramatické značkování, strojový překlad). Vychází sice z teorie FGP, ale zůstává přenositelný do jiných teoretických systémů. Základem je slovník BRIEF, jejž se pokouší vylepšit automatickým spojením rámců, které jsou oddělené, ač nesou stejný význam, jak jsme zmínili výše. Využívá také Slovesa pro praxi. Obsahuje zhruba 15 000 sloves.

Valenční rámec je tvořen původními funktory z FGP (tedy např. „Kam“ a „Jak dlouho“ namísto DIR3 a THL) a jejich realizacemi na povrchové rovině, jak je známe např. z VALLEXu. Také funktory jsou přiřazeny automaticky.

Na rozdíl od předchozích slovníků s výjimkou[25] PDT-Vallexu 2.0 zachycuje vybrané diateze. Je zde též postižena reflexivita.

Český syntaktický lexikon nebyl nikdy zveřejněn. Nevýhodou může také být automatické sloučení rámců a přiřazení funktorů členům valenčního rámce, které není dokonale přesné.

4.5 WordNet

Lexikální databáze WordNet (Fellbaum, 1998) tvoří síť slov[26] vzájemně propojených na základě významu. WordNet je tak lexikálním zdrojem na pomezí mezi slovníkem a tezaurem. Dnes již existuje mnoho lexikálních databází odvozených z původního princetonského WordNetu pro angličtinu. Na princetonském WordNetu ukážeme jeho obecné vlastnosti a potom představíme Český WordNet.

V anglickém WordNetu 3.0 jsou popisovány pouze autosémantické slovní druhy, konkrétně substantiva, adjektiva, slovesa a adverbia. Nejdůležitější z mnoha druhů relací je synonymie, která určuje rozklad množiny významů do tříd ekvivalence nazývaných *synsety*. Každý synset zastupuje jeden koncept. Má-li slovo více významů, budou patřit do různých synsetů. Synset obsahuje jedno či více vzájemně synonymních slov[27] a je základním elementem slovníku. Každý synset

[25] Ve VALLEXu budou diateze komplexně zpracovány ve verzi 3.0.

[26] Přesněji síť významů. Jednotlivé významy polysémních slov jsou rozlišovány pomocí čísel připojovaných s dvojtečkou na konec lemmatu (kterou v této práci až na výjimky neuvádíme). Navíc tento význam může být tvořen víceslovným výrazem.

[27] Jde o slova, která jsou alespoň v nějakém kontextu synonymní.

obsahuje glosu, většina synsetů také příkladové věty. Slovesa navíc obsahují povrchový slovesný rámec v jednoduché podobě, např. „Somebody —s something from somebody“ pro synset {*distinguish, separate, differentiate, secern, secernate, severalize, severalise, tell, tell apart*}, nebo „Somebody —s somebody INFINITIVE“ pro {*invite, bid*}. Ve WordNet je pouze 35 různých slovesných rámců.

Substantivní a slovesné synsety jsou ve WordNetu uspořádány hierarchicky pomocí další zásadní relace, a to *hyponymie/hyperonymie*.[28] V nejvyšším patře takto tvořené hierarchie jsou základní koncepty, z nichž jsou všechny ostatní synsety tranzitivně dosažitelné relací hyponymie/troponymie. Všechna substantiva mají v princetonském WordNet jediný základní koncept, kořen celého stromu,[29] synset {*entity*} s popisem „that which is perceived or known or inferred to have its own distinct existence (living or nonliving)“. Slovesa mají 559 základních konceptů, slovesnou část WordNetu tedy tvoří les s 559 stromy zakořeněnými v synsetech reprezentujících základní koncepty. Najdeme mezi nimi skutečně základní pojmy jako {*change, alter, modify*} a {*be active, move*}, ale také pojmy poměrně speciální, třeba v synsetech {*clear up, clear, light up, brighten*} či {*nod*}.

Adjektiva netvoří hierarchii, jsou uspořádána do antonymních dvojic, kolem kterých jsou sdružena adjektiva podobná. Například skupina podobných synsetů {*straight*}, {*correct, right*}, {*accurate, exact, precise*}, {*letter-perfect, word-perfect*} je přes dvojice antonym *correct—incorrect* a *right—wrong*[30] nepřímo antonymní se skupinou {*erroneous*}, {*incorrect, wrong*}, {*fallacious*}, {*false, mistaken*}.

Adverbií je ve WordNetu málo a *relační adverbia* jsou obvykle pouze propojena relací nazývanou ve WordNetu „*pertainym*“ s adjektivem, ze kterého jsou odvozena.

Další relace, které jsou ve WordNetu přítomny, netvoří ucelenou síť, propojují některé ze synsetů „napříč“ stromovou strukturou. Patří sem například nesymetrická relace *holonymum–meronymum* pro substantiva, či jednosměrný *entailment* pro slovesa, tedy činnost, která je slovesem nutně implikována, např. {*succeed*} → {*try*}.

[28] V případě sloves jde o *troponymii*/hyperonymii.

[29] Máme-li být přesní, musíme říci, že to není strom, nýbrž orientovaný graf, neboť sice výjimečně, ale přesto existuje v hierarchické struktuře vícenásobná dědičnost: například {*domestic cat, house cat, Felis domesticus, Felis catus*} má dvě hyperonyma: {*cat, true cat*} a {*domestic animal, domesticated animal*}. Tyto případy jsou však řídké, proto je pro představu o celkové struktuře stále výhodné si WordNet jako strom představovat.

[30] Většina relací je vedena mezi synsety, nicméně je možné spojovat také jednotlivé prvky ze synsetů, jako je tomu v případě etymologické příbuznosti slov nebo zde v případě antonym.

Až na výjimky (příbuzná slova) nejsou vedeny relace mezi různými slovními druhy, můžeme proto chápat WordNet jako čtyři oddělené lexikální databáze, které mají různou velikost i různou strukturu a vnitřní vztahy.

WordNet obsahuje přes 117 000 synsetů, většina z nich je substantivní.

4.5.1 Formát princetonského WordNetu

WordNet 3.0 je zdarma ke stažení[31] pro vědecké i komerční účely (s podmínkou správné citace původního projektu a zachování stejné licence i pro odvozená díla) spolu s některými nástroji pro práci s ním. Data jsou uložena v poměrně jednoduchém, nicméně proprietárním textovém formátu. Jeden řádek reprezentuje jeden synset a například pro synset {*cat, true cat*} vypadá řádek takto:

```
02121620 05 n 02 cat 0 true_cat 0 003
  @ 02120997 n 0000
  ~ 02121808 n 0000 ~ 02124623 n 0000
  | feline mammal usually having thick soft fur
    and no ability to roar: domestic cats; wildcats
```

První položka je unikátní číslo synsetu, dále vidíme například slovní druh (`n` = noun) a počet členů synsetu (`02`), za nimiž následuje jejich výčet s odlišením homonym (zde `0`). Za „zavináčem“ je odkaz na hyperonymum {*feline, felid*} a za tildou odkaz na hyponymum {*domestic cat, house cat, Felis domesticus, Felis catus*} a {*wildcat*}. Synset končí glosou uvozenou svislicí. Obdobným způsobem jsou kódovány i další relace zmíněné v sekci 4.5. Informace o slovesném rámci jsou například uloženy jako odkaz do tabulky na jeden ze zmíněných 35 použitých rámců.

4.5.2 Český WordNet

Česká verze lexikální databáze WordNet (Pala a Ševeček, 1999; Smrž, 2004) vznikla v Centru zpracování přirozeného jazyka na Fakultě informatiky Masarykovy univerzity v Brně v rámci projektů EuroWordNet (Vossen, 1998) a BalkaNet (Tufiş, Cristea a Stamou, 2004), kde byly vyvíjeny i další jazykové verze WordNetu. Ve všech jsou synsety (pomocí tzv. Inter-Lingual Indexu) namapované na synsety v princetonském WordNetu 2.0. Český WordNet vznikal poloautomaticky: překladem z princetonského i vyhledáváním konceptů v dostupných slovnících.

Verze 1.9 (Pala et al., 2010), která byla upravena v Ústavu formální a aplikované lingvistiky v Praze, obsahuje přes 23 000 synsetů. Byla použita k anotaci

[31] `https://wordnet.princeton.edu/wordnet/download`

všech autosémantických slov v PDT, která se v Českém WordNet vyskytovala (což tvoří 49,5 % všech autosémantik). Vychází ze stejně objemné verze databáze, ale v průběhu anotace zde byly upraveny a doplněny synsety vytipované jako nejproblematičtější. Projekt je popsán (včetně úpravy lexikální databáze) v Bejček, Möllerová a Straňák (2006).

Z důvodu poloautomatického vzniku slovníku není hierarchická struktura tvořená hypo-/hyperonymickou relací dostatečně propojená a Český WordNet obsahuje téměř tisíc základních slovesných a substantivních základních obecných konceptů, tedy kořenů jednotlivých stromů, které jsou navíc často zcela izolované a nemají ani žádná hyponyma. Je mezi nimi např. {*muka, utrpení, agónie*}, nebo {*dávit, blinkat, blít, poblinkat, pozvracet*}.

4.5.3 Formát Českého WordNetu

Český WordNet 1.9 je také volně dostupný[32] pod licencí CC BY-NC-SA 3.0 a je ve formátu XML. Informace jsou pochopitelně podobné jako u princetonského WordNetu, jen zapisované explicitněji.

```
<SYNSET><ID>01457160-n</ID><POS>n</POS>
  <SYNONYM><LITERAL>kočka<SENSE>1</SENSE></LITERAL></SYNONYM>
  <ILR>01456555-n<TYPE>hypernym</TYPE></ILR>
</SYNSET>
```

Synset je tvořen svým jednoznačným identifikátorem, slovním druhem a seznamem synonym – v tomto případě jediné, {*kočka:1*}. Následuje odkaz na hyperonymum {*kočkovitý, kočkovitá šelma*}. V Českém WordNet vedou tyto odkazy vždy pouze na nadřazený pojem, naopak od hyponym jako jsou *kocour*, *mourek*, *perská kočka* a mnoho dalších vede hyperonymický odkaz na tento synset `01457160-n`.

4.6 FrameNet

Projekt FrameNet vychází z Fillmorovy teorie *rámcové sémantiky* (*frame semantics*) a jeho počátky tedy sahají do sedmdesátých let (Fillmore, 1976, 1982). Zabývá se především slovesy, vybranými substantivy a adjektivy a hlavně zachycením situací, které tato slova popisují.

FrameNet je vedle projektu také název jeho výsledku, lexikální databáze[33] (Ruppenhofer et al., 2006), podobně jako WordNet. Tvoří ji **graf**, jehož uzly (*sémantické rámce*) reprezentují nějakou situaci, objekt, či událost a teprve jim

[32] `http://hdl.handle.net/11858/00-097C-0000-0001-4880-3`

[33] FrameNet je dostupný online i ke stažení na `https://framenet.icsi.berkeley.edu`.

jsou přiřazeny *lexikální jednotky*, tedy slova v určitém významu (typicky slovesa či dějová substantiva). Tyto rámce jsou propojeny do obrovské sítě pomocí osmi relací. Pro určitou představu o celkové struktuře FrameNetu by měl posloužit též obrázek 4.3, který zachycuje malý výsek této sítě. Nyní si databázi popíšeme podrobněji.

FrameNet popisuje všechny přípustné syntaktické a sémantické vazebné možnosti slov (tedy jejich valenci), a to ve všech významech, kterých mohou slova nabývat. Jednu situaci, jakou je například narození, zachycuje v jednom **sémantickém rámci**, který tzv. *„evokují“* příslušná slova v tomto významu, tedy lexikální jednotky (ve stejném významu jako jsme pojem používali ve VALLEXu). Je-li slovo homonymní, jednotlivé jeho významy evokují různé rámce. Situace může být nahlížena i z více perspektiv, proto je sémantický rámec Birth_scenario evokován[34] například těmito slovy (vždy v příslušném významu): *born.v, come into world.v, birth.n, spawn.v, calving.n* a mnoho dalších (písmeno za tečkou ve FrameNetu značí slovní druh). Podle autorů FrameNetu chápou uživatelé jazyka významy slov na základě sémantických rámců.

Sémantický rámec obsahuje svou definici či glosu[35] a dále ***elementy rámce*** (FE, *frame elements*). Jsou to například účastníci děje, či jeho průvodní okolnosti, v příkladu v poznámce pod čarou jsou všechny FE vyznačeny kapitálkami. Elementy rámce jsou společné pro všechny lexikální jednotky rámce. Dělí se zejména na elementy tvořící jádro („core“) a rámce periferní („non-core, peripheral“):[36] v jádru jsou ty elementy, které jsou pro daný rámec koncepčně nezbytné; jejich výčet zároveň odlišuje jeden rámec od ostatních. Periferní elementy nemohou být subjektem ani objektem slovesa, mohou se vyskytovat téměř ve všech rámcích a obvykle jsou tvořeny adverbii či předložkovými frázemi.

Některé elementy rámce mohou mít dále přiřazeny další informace, zejména sémantický typ, což jsou například hodnoty Sentient pro roli Agent, Locative_relation pro Place, nebo Human pro Examiner. Dále to jsou omezení (exclude a require), kdy jeden rámec vylučuje, nebo naopak vyžaduje přítomnost rámce jiného, čehož se využívá například pro řešení recipročních konstrukcí („Item-1 *conflicts with* Item-2“/„Items *conflict*“) či lexikálních alternací („Agent *ties* Item *to* Goal“/„Agent *ties* Items“) a další.

[34] Přesněji tento rámec pomocí relace Perspective On (viz dále) obsahuje dva další rámce, Being_born a Giving_birth (viz obrázek 4.3 vlevo) a teprve tyto dva rámce jsou evokovány zmíněnými slovy.

[35] Např. *„A* Mother *and* Father *produce a* Child *or an* Egg.“ pro sémantický rámec Giving_birth nebo *„An* Offspring *is born to a* Mother *and* Father*, collectively referred to as* Parents*. This event happens at a particular* Time *and* Place.“ pro Birth_scenario.

[36] Další dvě hodnoty jsou „extra-thematic“ pro popis např. dějů na pozadí a „core-unexpressed“ pro případy kontroly.

Díky tisícům anotovaných příkladových vět (viz níže) je FrameNet s FE užitečným zdrojem dat pro úlohu přiřazování sémantických rolí (*semantic role labeling*).

Uvedeme příklad sémantického rámce Giving_birth (vybíráme jen důležitější informace, o kterých jsme zde mluvili):

Definition:

A MOTHER and FATHER produce a CHILD or an EGG.
„Betty BORE Gerry three intelligent daughters.“

Core Frame Elements:

CHILD	CHILD identifies the new self-motile creature produced from the MOTHER and FATHER. *„Betty BORE Gerry* three intelligent daughters.“
EGG	EGG is an immobile object containing an organism that
Excludes:	may hatch as a mobile, infant organism.
CHILD	*„Female clownfish LAY* their eggs *around sea anemones.“*
FATHER	FATHER is the male creature that copulates with the MOTHER, thus leading to the birth of the CHILD. *„Betty BORE* Gerry *three intelligent daughters.“*
MOTHER	MOTHER is the female creature that produces the CHILD. „Betty *BORE Gerry three intelligent daughters.“*
PARENTS	The MOTHER and FATHER expressed together. „We *are going to be HAVING twins!“*

Non-Core Frame Elements:

CIRCUMSTANCES	This FE identifies the CIRCUMSTANCES under which a MOTHER and FATHER produce produce a CHILD. *„Star fish can be SPAWNED* under the right conditions.“
DEPICTIVE	The state of the CHIILDE as it enters the world. *„Demons are SPAWNED* fully formed.“

MANNER Semantic Type: Manner	Any description of the intentional act which is not covered by more specific FEs, including secondary effects (quietly, loudly), and general descriptions comparing events (the same way). In addition, it may indicate salient characteristics of the PARENTS (or a FATHER or MOTHER) that also affect the action (presumptuously, coldly, deliberately, eagerly, carefully). *„They* eagerly *SPAWNED more of their kind until there was no room for more.“*
MEANS Semantic Type: State_of_affairs	This FE identifies the MEANS by which the event occurs.
PLACE Semantic Type: Locative_relation	This FE identifies the PLACE where the event occurs.
RESULT	This FE identifies the RESULT of the birth. *„John's mother BORE him* into a wealthy family.“
TIME Semantic Type: Time	This FE identifies the TIME when the event occurs.

Frame-frame Relations:
Perspective on: Birth_scenario
Precedes: Death, Dying

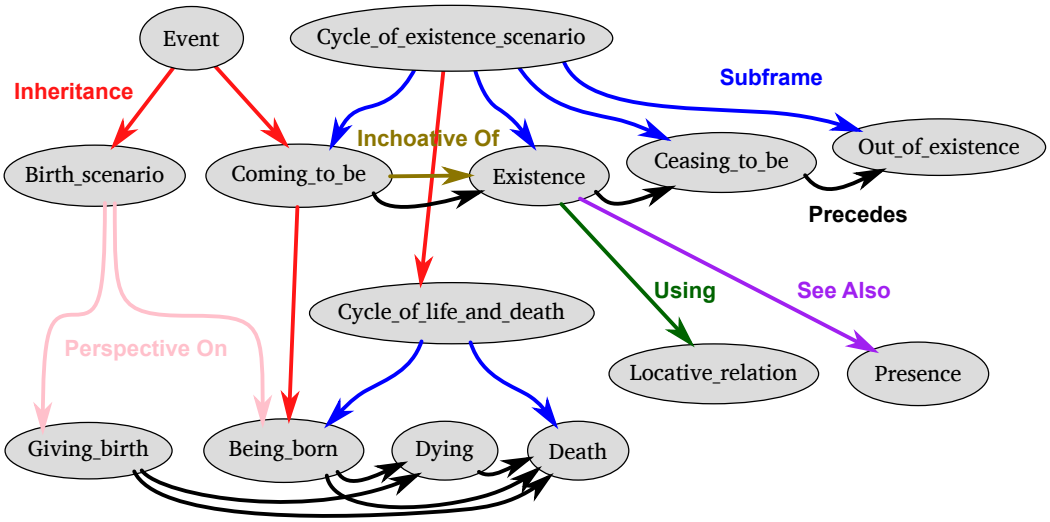
Lexical Units:
bear.v, beget.v, birth.n, birth.v, bring forth.v, calve.v, calving.n, carry to term.v, drop.v, father.v, generate.v, get.v, have.v, kid.v, lambing.n, lay.v, mother.v, propagate.v, sire.v, spawn.v, whelp.v

Zmínili jsme v úvodu této sekce, že sémantické rámce jsou propojeny bohatou sítí **relací**. V příkladu vidíme dvě: Perspective On a Precedes. Celkem je ve FrameNetu definováno osm druhů těchto relací, z nichž sedm je vidět na obrázku 4.3.

Inheritance Nejdůležitější relace dědičnosti vlastně odpovídá relaci „is-a“ a vede ke specifičtějším rámcům, pro které platí vše co pro rodiče.

Subframe Komplexní rámce, sestávající z více menších situací, které spolu úzce souvisí či na sebe navazují, byly rozděleny na podrámce a propojeny relací Subframe.

↪ *Příklad:* Rámec Cycle_of_life_and_death sestává ze tří rámců: Being_born, Death a z obrázku vynechaného Dead_or_alive.



Obrázek 4.3: Ukázka relací mezi sémantickými rámci FrameNetu. Jsou vybrány rámce týkající se změny existence a všechny relace mezi nimi jsou zachovány (ovšem desítky dalších relací k rámcům, které nebyly do ukázky vybrány, zde nejsou ani naznačeny – uvádíme tedy indukovaný podgraf). Sedm druhů relací je barevně odlišeno a pojmenováno.

Precedes V takovém případě mohou být podrámce ještě mezi sebou propojeny relací Precedes, mají-li pevně danou posloupnost, v jaké nastávají.

↪ *Příklad:* Being_born i Giving_birth předcházejí rámci Dying, a ten zas předchází rámci Death.

Using Relace Using je jakousi náhradou za dědičnost, pokud tato nelze použít (například z důvodu vícenásobné dědičnosti).

↪ *Příklad:* Tak třeba rámec Judgment_communication (mj. *blame.v, critic.n, derision.n, praise.v*) není potomkem ani Judgment (*admire.v, disapprove.v*), ani Statement (*explain.v, insist.v*), ale z obou přebírá mnohé FE, proto od obou vede relace Using.

Perspective On V případě nejméně dvou možných úhlů pohledu na danou situaci jsou tyto specifičtější, ale různé situace spojeny s obecnou situací relací Perspective On.

↪ *Příklad:* Od Commerce_goods-transfer vede relace ke Commerce_buy (evokovanému např. slovy *buyer.n, purchase.v*) a ke Commerce_sell (*vendor.n, retail.v, sale.n*).

Inchoative Of Rámce vyjadřující změnu stavu jsou touto relací spojeny s rámci stavovými, které představují výsledek oné změny stavu.

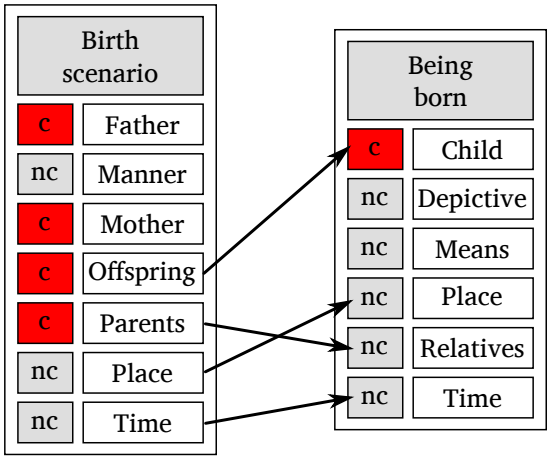
↪ *Příklad:* Do sémantického rámce Position_on_a_scale vede relace Inchoative Of z rámce Change_position_on_a_scale. Na našem obrázku to jsou rámce Coming_to_be a Existence.

Causative Of Rámce, které vyjadřují činnost způsobující změnu stavu, jsou spojeny s rámci reprezentujícími tuto změnu relací Causative Of.

↪ *Příklad:* Ve zmíněném rámci Change_position_on_a_scale končí relace Causative Of vycházející z rámce Cause_change_of_scalar_position.

See Also Poslední relace není ničím specifická, používá se pro porovnání příbuzných rámců. Mohla-li by nastat nejistota, v čem se skupina podobných rámců liší, je vybrán jeden z nich, v jeho definici jsou potom uvedeny odlišnosti od zbylých rámců a ty na něj odkazují relací See Also.

↪ *Příklad:* Rámec Existence na obrázku 4.3 (definice „An Entity is declared to exist, generally irrespective of its position or even the possibility of its position being specified.“) odkazuje na rámec Presence, kde je kromě definice („An Entity exists at a particular Location, at a particular Time, as observed by an implicit observer.“) uvedena také odlišnost od Existence: „This frame differs from Existence in that the Location is profiled as a ground where an observer is conceived of as confirming the Entity's existence.“



Obrázek 4.4: Ukázka relace Perspective On mezi dvěma sémantickými rámci FrameNetu. Vztah mezi oběma rámci je rozepsán na jednotlivé elementy rámce a ty, které v obecném rámci Birth_scenario odpovídají některým ze specifičtějšího rámce Being_born, jsou vyznačeny šipkami. „Core“ elementy jsou vyznačeny červeně.

V případě, že jsou dva rámce propojeny některou ze zmíněných relací, jsou také propojeny ty elementy rámce, které si vzájemně odpovídají, jak je vidět na obrázku 4.4 pro relaci Perspective On z levé části předchozího obrázku. Takto korespondujících elementů může být víc než deset v jedné relaci.

FrameNet byl použit také k rozsáhlé **anotaci** vět z korpusu, díky nimž většina rámců nabízí příkladové věty. Většina vět byla vytipována pro dané slovo, anotátor se potom v celé větě zaměřil právě na toto slovo: přiřadil mu sémantický rámec (a tím zároveň určil, o jakou LU se jedná) a zbylým členům věty přiřadil patřičné FE, dále syntaktickou roli (větný člen) a fráze z frázové gramatiky. (Zlomek vět byl naopak anotován vícenásobně a sémantický rámec se přiřazoval postupně všem slovům, pro které existoval.)

Autoři tvrdí, že takto získané příkladové věty ilustrují všechny možné kombinace dané lexikální jednotky (Ruppenhofer et al., 2006, str. 6). Díky tomu mohly být anotace pro každou LU převedeny na seznam všech přípustných povrchových valenčních kombinací, které jsou ve FrameNetu i s jejich frekvencí a odkazy na konkrétní věty také k disposici.[37]

FrameNet verze 1.5 obsahuje téměř 12 000 lexikálních jednotek (jejich přesné rozložení podle slovních druhů podává tabulka 4.2), které jsou evokované více než tisícem sémantických rámců. Vše doplňuje více než 175 000 korpusových dokladů.

Vedle původního anglického FrameNetu již dnes vzniká celá řada jazykových variant, mj. španělský, německý, čínský, dánský či japonský FrameNet.

[37] Začínali jsme příkladem s narozením. Abychom se přidrželi rámce Cycle_of_life_and_death, uveďme příklad z rámce Death, konkrétně všech šest valenčních kombinací pro sloveso *pass away*. Za elementem rámce uvádíme v závorce gramatickou funkci a za lomítkem typ fráze. Pro bližší výklad viz Ruppenhofer et al. (2006). Každou valenční kombinaci doplňujeme jednou anotovanou větou z FrameNetu.

1. Explanation(PP[with]/NP) Protagonist(NP/Ext)
   Me own mammy PASSED AWAY with the consumption (…).
2. Explanation(PP[because of]/Dep) Protagonist(NP/Ext) Time(PP[in]/Dep)
   He sadly PASSED AWAY prematurely in 1962 because of severe cerebrovascular disease.
3. Protagonist(NP/Ext) Time(NP/Dep)
   A living legend PASSED AWAY when Ferdinando Keast died in 1891.
4. Protagonist(NP/Ext) Time(Sinterrog/Dep)
   Your father PASSED AWAY about four minutes ago.
5. Protagonist(NP/Ext)
   Hi my name is Sandra Reid (…) and my mom and dad have PASSED AWAY.
6. Manner(AVP/Dep) Place(PP[in]/Dep) Protagonist(NP/Ext)
   (…) but he PASSED AWAY peacefully in his bed.

Zdůrazněme, že z hlediska elementů rámce jde pouze o pět kombinací, neboť třetí a čtvrtý případ se liší gramatickou funkcí. I toto je ve FrameNetu odlišeno, kombinace FE se dále dělí na kombinace specifičtější.

Slovní druh	Počet LU
substantiva	4 743
slovesa	4 605
adjektiva	2 122
adverbia	167
předložky	143
číslovky	33
spojky	6
členy	4
podřadící spojky	3
citoslovce	3
Celkem	11 829

Tabulka 4.2: Zastoupení slovních druhů lexikálních jednotek ve FrameNetu. Víceznačné slovo sestává z více lexikálních jednotek v různých rámcích a každá je zde započítána zvlášť. (Například samotné sloveso *take* – nepočítáme-li dalších 18 frází a frázových sloves – evokuje osm různých sémantických rámců.)

4.6.1 Formát FrameNetu

FrameNet je distribuován jako tisíce XML souborů (zejména 1 000 pro rámce, 12 000 pro lexikální jednotky) spolu s XSLT šablonami, které umožňují jejich transformaci do HTML, a tudíž bezproblémové prohlížení ve webovém prohlížeči, který tyto transformace umí provádět transparentně za běhu. Výsledek je shodný s webem FrameNetu, ovšem bez nástroje FrameGrapher, který umožňuje zobrazovat vztahy mezi sémantickými rámci podobné obrázku 4.3.

Následující velmi zkrácená ukázka reprezentuje v XML formátu některé informace o rámci Giving_birth, které jsme viděli v příkladu na straně 42.[38]

Rámec Giving_birth v XML formátu FrameNetu

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<?xml-stylesheet type="text/xsl" href="frame.xsl"?>
<frame name="Giving_birth" ID="992874" …>
  <definition>
    [def-root]A [fen]Mother[/fen] and [fen]Father[/fen] produce a [fen]Child[/fen] or an
      [fen]Egg[/fen].
      [ex] [fex name="Mother"]Betty[/fex] [t]bore[/t] [fex name="Father"]Gerry[/fex]
```

[38] Protože původní XML, které obsahuje vnořené části také jako XML, je špatně čitelné, přepíšeme vnitřní XML tagy hranatými závorkami. Místo `<definition>This FE identifies the &lt;fen&gt;Time&lt;/fen&gt;.&lt;</definition>` budeme psát `<definition>This FE identifies the [fen]Time[/fen].</definition>`.

```
        [fex name="Child"]three intelligent daughters[/fex].[/ex] [/def-root]
    </definition>
10  <FE Color="800080" coreType="Core" name="Child" ID="995620" …>
      <definition>
        [def-root][fen]Child[/fen] identifies the new self-motile creature produced from the
          [fen]Mother[/fen] and [fen]Father[/fen].
          [ex]Betty [t]bore[/t] Gerry [fex name="Child"]three intelligent
15          daughters[/fex].[/ex][/def-root]
      </definition>
    </FE>
    …
    <FE Color="9400D3" coreType="Peripheral" name="Depictive" ID="995624" …>
20    <definition>
        [def-root]The state of the [fen]Child[/fen] as it enters the world.
          [ex]Demons are [t]spawned[/t] [fex name="Depictive"]fully formed[/fex].[/def-root]
      </definition>
    </FE>
25  …
    <frameRelation type="Is Preceded by"/>
    <frameRelation type="Perspective on">
        <relatedFrame>Birth_scenario</relatedFrame>
    </frameRelation>
30  <frameRelation type="Precedes">
        <relatedFrame>Death</relatedFrame>
        <relatedFrame>Dying</relatedFrame>
    </frameRelation>
    <frameRelation type="Inherits from"/>
35  …
    <lexUnit POS="V" name="carry to term.v" ID="999790" …>
      <definition>FN: be pregnant with and give birth to</definition>
      <sentenceCount annotated="0" total="0"/>
      <lexeme order="1" headword="false" breakBefore="false" POS="V" name="carry"/>
40    <lexeme order="2" headword="false" breakBefore="false" POS="PREP" name="to"/>
      <lexeme order="3" headword="false" breakBefore="false" POS="N" name="term"/>
    </lexUnit>
    <lexUnit POS="N" name="birth.n" ID="999791" …>
      <definition>COD: give birth to</definition>
45    <sentenceCount annotated="2" total="2"/>
      <lexeme order="1" headword="false" breakBefore="false" POS="N" name="birth"/>
    </lexUnit>
    …
</frame>
```

XML soubor jsme zobrazili od prvního po poslední řádek, ovšem obdobné, opakující se části uprostřed jsme vypustili. Na 3. řádku rámec začíná, obsahuje jméno, identifikační číslo a další vypuštěné údaje. Od řádku 4 do řádku 9 vidíme zapsanou definici rámce a příklad. Následuje výčet elementů rámce (10–25); ty obsahují opět definici a případně příklad. Některé jsou „core“ (10), jiné jsou periferní (19). Od řádku 26 jsou uváděny vztahy k ostatním rámcům FrameNetu. Na závěr přichází výčet lexikálních jednotek náležejících danému rámci, my zde uvádíme jen dvě: *carry to term.v* (od ř. 36) a *birht.n* (od ř. 43). Obě obsahují počet

anotovaných příkladů, které jsou uloženy v jiném XML souboru (odkazovaném přes ID této LU); jeho formát je podobně srozumitelný[39] a ukázku již neuvádíme.

4.7 Další anglické slovníky a lexikální databáze

Vedle WordNetu a FrameNetu představíme ještě další lexikální databáze pro angličtinu, PropBank a VerbNet, které jsou společně provázány v projektu SemLink (sekce 4.8), a databázi NomBank pro substantiva.

4.7.1 PropBank

Proposition Bank I, zkráceně nazývaný PropBank,[40] byl projekt, v rámci něhož byla provedena základní sémantická anotace (vztahy mezi predikátem věty a jeho argumenty) na anglických datech v Penn Treebanku[41] (konkrétně v sekci Wall Street Journalu, WSJ, která obsahuje více než milion slov) (Palmer, Kingsbury a Gildea, 2005). Od počátku bylo cílem připravit trénovací data pro *semantic role labeling*. Anotováno bylo kolem 133 000 sloves, což jsou veškerá slovesa v korpusu s výjimkou slovesa *be* a dále sloves *do* a *have* užitých jako pomocná.

Součástí dat vydaných v roce 2004 v LDC[42] je kromě zmíněných anotací také slovník sloves čítající přes 3 200 hesel. Každé heslo obsahuje seznam *rolí*, neboli argumentů slovesa. Role jsou číslovány a doplněny o stručnou sémantickou charakteristiku pro rozlišení.

↪ *Příklad:*

tolerate `Arg0:` *tolerater* `Arg1:` *thing tolerated*

whirl `Arg1:` *thing in motion* `Arg2:` *axis (rare)*

buy `Arg0:` *buyer* `Arg1:` *thing bought* `Arg2:` *seller* `Arg3:` *price paid* `Arg4:` *benefactive*

Po výčtu rolí následují příklady jejich možných kombinací, což jsou v zásadě návody pro anotaci. Uvedeme příklad celého slovesa *think*:

---

[39] XML pro LU recipročně obsahuje informaci o sémantickém rámci, kam daná jednotka patří, tedy odkaz zpět na ukázané XML, dále o elementech rámce a pak následuje už jen řada vět, ve kterých jsou tři vrstvy anotace: elementy rámce, gramatická funkce a typ fráze. (Zajímavý se může zdát způsob, jakým je určován rozsah značek ve větě. Formát jednoduše využívá implicitní číslovaní znaků od začátku věty a značka pak má rozsah např. od 143. do 161. písmene.)

[40] `http://verbs.colorado.edu/~mpalmer/projects/ace.html`

[41] `https://www.cis.upenn.edu/~treebank`

[42] `https://catalog.ldc.upenn.edu/LDC2004T14`

Predicate: think
Roleset id: think.01
Roles:
`Arg0`: Thinker
`Arg1`: Thought
`Arg2`: Attributive (please look at the examples – this is tricky)
Example: transitive
A Lorillard spokeswoman thought "This is an old story."
`Arg0`: A Lorillard spokeswoman
`Rel`: thought
`Arg1`: "This is an old story"
Example: intransitive, with subject of thinking
People weren't thinking about targeting 10 years ago.
`Arg0`: People
`Argm-neg`: n't
`Rel`: thinking
`Arg2`: about targeting
`Argm-tmp`: 10 years ago
Example: transitive, with subject of thinking
People weren't thinking anything about targeting 10 years ago.
`Arg0`: People
`Argm-neg`: n't
`Rel`: thinking
`Arg1`: anything
`Arg2`: about targeting
`Argm-tmp`: 10 years ago
Example: attributive
They think of us as a good partner.
`Arg0`: They
`Rel`: think
`Arg1`: of us
`Arg2`: as a good partner
(Následuje ještě několik příkladů, které řeší problém s tzv. *traces*,[43] jako například *"What you have to understand," thought [*?*] John, "is that Philly literally stinks."*, neboť PropBank mj. vyžaduje, aby argumenty slovesa byly syntaktickými složkami. Tyto příklady zde vypouštíme.)

[43] Viz poznámku pod čarou č. 5 na straně 14.

Predicate: think_up
Roleset id: think.02
Roles:
`Arg0`: thinker
`Arg1`: thing thought up
Example: …

Predicate: think_over
Roleset id: think.03
Roles:
`Arg0`: thinker
`Arg1`: subject matter considered
Example: …

Predicate: think_through
Roleset id: think.04
Roles:
`Arg0`: thinker
`Arg1`: topic considered fully
Example: …

Soubory s jednotlivými lemmaty slovníku jsou uloženy ve vlastním formátu jako XML, který nevychází z lexikografických standardů, jsou však srozumitelné a čitelné. Anotace Penn Treebanku je distribuována odděleně od dat treebanku („stand-off“) v poměrně složitém textovém formátu, kde jeden řádek odpovídá jednomu anotovanému slovesu a jsou tam v osmi a více sloupcích zakódovány veškeré informace:

```
wsj/00/wsj_0067.mrg   4   30   gold
      think.01   vn--a
      24:1*28:1*29:0-ARG0   30:0-rel   31:1-ARG2-of   33:2-ARG1-as
```

Projekt PropBanku nadále pokračuje, rozšiřuje se množství anotovaných slov, anotují se jiné domény, vznikají paralelní vícejazyčné PropBanky, dochází k propojování s dalšími lexikálními databázemi (viz sekce 4.8) a k anotaci dalších slovních druhů (adjektiv, substantiv i víceslovných analytických predikátů) (Bonial et al., 2014).

4.7.2 NomBank

Projekt NomBank[44] je spjatý a koordinovaný s PropBankem – na stejných datech (Penn Treebank, část WSJ) anotuje argumenty substantiv.

[44] `http://nlp.cs.nyu.edu/meyers/NomBank.html`

NomBank 1.0 (Meyers et al., 2004) obsahuje zhruba polovinu substantiv vycházejících ze sloves (kromě nominalizace jde také o vztahy jako *agrese—zničit*), proto autoři vycházeli z obdobných argumentů příslušných sloves v PropBanku. Mezi zbylými substantivy jsou adjektivní nominalizace (např. *ability*), vztahová substantiva (např. *father*), partitiva (např. *handful*) a další.

Ze substantiv z WSJ jsou vybrána pouze ta, která někdy vážou argument, a více než polovina jejich výskytů, 114 576 instancí, byla anotována, neboť se vyskytla ve větě i se svými argumenty. Slovník pak obsahuje 4 700 substantiv – v jednom až čtrnácti významech.

XML formát souborů slovníku je totožný s formátem PropBanku.

4.7.3 VerbNet

Projekt VerbNet[45] (Schuler, 2006) kombinuje PropBank a sémantické třídy Levinové (Levin, 1993). Slovesa z PropBanku slučuje do hierarchicky uspořádaných sémantických tříd, které vycházejí z hierarchie Levinové, ale dále je rozšiřují a prohlubují.

VerbNet 3.2 obsahuje 4 400 sloves ve více než 6 300 významech rozdělených do 273 tříd a 214 podtříd. V nich je použito 30 tématických rolí.

Na obrázku 4.5 uvádíme opět příklad slovesa *think*. Spadá (spolu s dalšími 19 lemmaty) do třídy Consider-29.9, do podtřídy 29.9-2.[46] Podtřída 29.9-2 pro svých šest lemmat dědí tři *tématické role* Agent, Theme a Attribute ze třídy 29.9 a rozpracovává jejich možné kombinace, lexikální a sémantická omezení apod. Je důležité dodat, že Consider-29.9-2 není jediná třída, ve které se *think* vyskytuje. V druhém svém významu patří do třídy Wish-62 spolu se slovesy *expect, dream, hope, imagine* apod.

Data jsou rozdělena do souborů podle jednotlivých tříd a jejich XML formát je opět vytvořený přímo pro potřeby VerbNetu a na první pohled srozumitelný.

4.8 SemLink

Slovníky a lexikální databáze podobného zaměření je výhodné vzájemně prolinkovávat, a využívat tak pohromadě pro jedno sloveso informací ze všech slovníků naráz.

[45] `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`

[46] Takto jemné podtřídy mimochodem Levinová nezavádí. Ale i tak u ní sloveso *think* spadá do deklarativní třídy 29.4.

Obrázek 4.5: Jedno ze dvou sloves *think* ve VerbNetu, tentokrát v sémantické třídě Consider-29.9-2. Většina podtříd z této třídy je na obrázku vynechána, hierarchie je zobrazena v pravém horním rohu.

SemLink[47] je název projektu propojujícího pomocí vzájemných linků následující čtyři lexikografické zdroje, které jsme představili na předchozích stránkách:

- PropBank (4.7.1)
- VerbNet (4.7.3)
- FrameNet (4.6)
- WordNet (4.5)[48]

Tyto slovníky a databáze jsou propojeny nad daty WSJ. Tatáž data, která byla anotována pouze PropBankem, jsou tedy nyní anotována pomocí všech čtyř lexikografických zdrojů. Vedle této anotace je také možné získat soubory, které nabízejí provázání samotných slovníků – a to vždy po dvojicích. Případně využít *Unified Verb Index*[49] (UVI), který umožňuje procházení 8 537 propojenými slovesy online.

Několikrát už jsme pracovali se slovesem *think*, UVI o tomto slovese uvádí (a umožňuje na webu také „prokliknout“ do uvedeného zdroje) následující:

think:

Consider-29.9-2, Wish-62
*(odkazy na dvě třídy **VerbNetu**; první třídu jsme viděli na obrázku 4.5)*

(PropBank)
*(odkaz do **PropBanku** na čtyři slovesa think, think_up, think_over a think_through, která jsme viděli na straně 49)*

(fn Cogitation), (fn Opinion), (fn Regard), (fn Awareness),
*(odkazy na čtyři sémantické rámce **FrameNetu**, z nichž každý obsahuje mezi svými lexikálními jednotkami think)*

(Grouping)
*(odkaz na stránku se šesti významy pospojovanými z **WordNetu**, o nichž jsme hovořili v poznámce pod čarou číslo 48)*

Nyní se krátce podíváme na data, která je možné si stáhnout, tedy jak jsme zmínili, na anotaci WSJ a na propojení dvojic slovníků. „Stand-off“ anotace WSJ,

[47] `https://verbs.colorado.edu/semlink`, kde je i ke stažení

[48] Respektive tzv. OntoNotes Sense Groupings, `http://verbs.colorado.edu/VSAP/about.html`, které z WordNet vycházejí. Projekt reagoval na problém, že významy ve WordNet jsou členěny zbytečně jemně, což snižuje kvalitu všech výsledků (automatických, ale i shodu anotátorů). Výsledkem jsou hruběji členěné významy ze synsetů ve WordNet (Pradhan et al., 2007).
↪ *Příklad:* Sloveso *think* se vyskytuje ve WordNetu ve 13 synsetech, v OntoNotes Sense Groupings má jen pět významů (odpovídajících ve WordNetu významům 1+2+10+11, 3+8+9+12+13, 4, 5+6+13 a 7 – třináctý význam se opakuje ve dvou nových skupinách) plus šestý význam pro víceslovná použití.

[49] `http://verbs.colorado.edu/verb-index/index.php`

kterou používá SemLink stejně jako PropBank, je uložená ve formátu, který přímočaře rozšiřuje formát používaný v PropBanku (str. 51):

```
wsj/00/wsj_0067.mrg   4   30   gold
     think-v   29.9-2   IN   think.01   null   vn--a
     24:1*28:1*29:0-ARG0=Agent   30:0-rel
     31:1-ARG2=Attribute   33:2-ARG1=Theme
```

První řádek je totožný, druhý obsahuje lemma, třídu z VerbNetu, rámec z FrameNetu (v tomto případe „indefinite“), „roleset“ z PropBanku, číslo skupiny z OntoNotes Groupings (v tomto případě žádná neodpovídala) a opět morfologickou informaci. Zbylé řádky jsou opět stejné, jen se vedle PropBankových argumentů objevují navíc tématické role z VerbNetu. V souboru je celý tento údaj na jediném řádku a jeden soubor obsahuje anotaci celého WSJ – opět shodně s PropBankem.

Naopak bilaterální vazby mezi dvojicemi slovníků jsou uloženy ve velice čitelném XML. Uvádíme tři ukázky:

Příklad 1: Význam z VerbNetu – význam z FrameNetu:

```
<vncls class="29.9-2" vnmember="think" fnframe="Awareness"
  fnlexent="169" versionID="vn1.5"/>
<vncls class="29.9-2" vnmember="think" fnframe="Opinion"
  fnlexent="" versionID="vn3.2"/>
<vncls class="29.9-2" vnmember="think" fnframe="Regard"
  fnlexent="" versionID="vn3.2"/>
```

Příklad 2: Tématické role z VerbNetu – element rámce z FrameNetu:

```
<vncls class="29.9" fnframe="Awareness">
  <roles>
    <role fnrole="Cognizer" vnrole="Agent"/>
    <role fnrole="Content" vnrole="Theme"/>
  </roles>
</vncls>
```

Příklad 3: „Roleset“ i argumenty PropBanku – třídy a tématické role VerbNetu:

```
<predicate lemma="think">
  <argmap pb-roleset="think.01" vn-class="29.9-2">
    <role pb-arg="0" vn-theta="Agent" />
    <role pb-arg="1" vn-theta="Theme" />
    <role pb-arg="2" vn-theta="Attribute" />
  </argmap>
  <argmap pb-roleset="think.01" vn-class="62">
    <role pb-arg="0" vn-theta="Experiencer" />
    <role pb-arg="1" vn-theta="Theme" />
  </argmap>
</predicate>
```

Okrajově zmiňme ještě projekt WordFrameNet[50] propojující WordNet s FrameNetem, který je zdarma ke stažení pod licencí CC BY 3.0, a jeho nástupce Predicate Matrix[51] (de Lacalle et al., 2014), který je k dispozici pod stejnou licencí. Propojuje tytéž čtyři zdroje, neboť z projektu SemLink vychází a využívá ho.

4.9 SemLex

Nyní přejdeme k odlišnému lexikografickému zdroji, ke slovníku víceslovných výrazů (VV). SemLex byl sestavený v Ústavu formální a aplikované lingvistiky v Praze během projektu Lexemann[52] (projekt stručně představujeme v sekci 6.1.3) a celý slovník je volně ke stažení pod licencí CC BY 3.0.

SemLex je tvořen 8 797 slovníkovými hesly, každé reprezentuje jeden VV. Slovník obsahuje všechny VV, které splňovaly stanovená kritéria[53] a byly nalezeny v korpusu PDT 2.0 během jeho anotace (Straňák, 2010). Základ slovníku (ještě před samotnou anotací PDT) vznikl spojením tří existujících slovníků: Slovníku české frazeologie a idiomatiky (Čermák, Červená, Churavý a Machač, 1994) a vybraných víceslovných výrazů z EuroVocu[54] a Českého WordNetu (Smrž, 2004) (více viz Straňák, 2010). Většina takto získaných VV se během anotace nikdy nepoužila a naopak mnohé další VV byly do slovníku přidány. Ve zveřejněném

[50] `http://adimen.si.ehu.es/web/WordFrameNet`

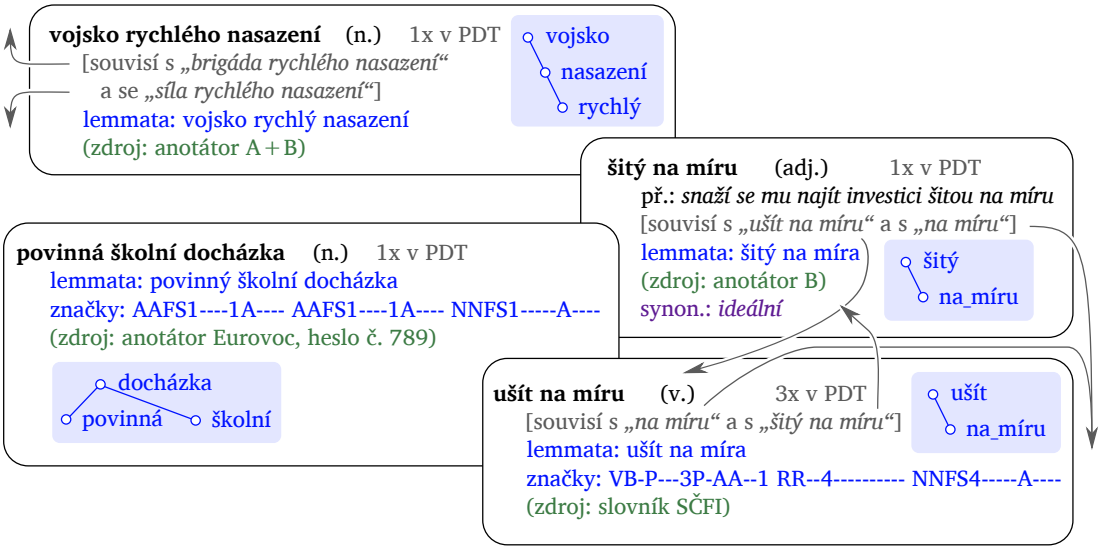
[51] `http://adimen.si.ehu.es/web/PredicateMatrix`

[52] `http://ufal.mff.cuni.cz/lexemann`

[53] Zjednodušeně lze říci, že hlavním kritériem byla sémantická nekompozicionalita výrazu. Více v sekci 6.1.2.

[54] EuroVoc je vícejazyčný polytematický tezaurus s terminologií z oblastí, kterými se zabývá Evropská unie; je přístupný ve 23 úředních jazycích EU z `http://europa.eu/eurovoc/`. Lze ho také stáhnout ve vlastním jednoduchém XML formátu.

SemLexu jsou pouze ta hesla, která se objevila v PDT. Pokud ovšem takové heslo pochází původně ze zdrojových slovníků, je u takového hesla uveden odkaz na heslo v původním slovníku a případně i další informace (například seznam synonym; glosa, příklad; přes původní identifikátor EuroVocu nebo SČFI je zase možné získat překlad do jiného evropského jazyka).



Obrázek 4.6: Vizualizace čtyř slovníkových hesel ve slovníku SemLex pro víceslovné výrazy *vojsko rychlého nasazení*, *šitý na míru*, *povinná školní docházka* a *ušít na míru*. Každé heslo obsahuje základní tvar (tučně), slovní druh a frekvenci výskytu v PDT. Dále seznam lemmat, ze kterých je heslo tvořeno (modře), tektogramatickou syntaktickou strukturu podstromu tvořícího výraz (rámeček) a zdroj, jak se heslo do slovníku dostalo (zeleně). Některá hesla SemLexu jsou mezi sebou propojována odkazy, více v sekci 7.2.2. V ukázce jsou odkazy do nezobrazených částí slovníku jen naznačeny.

Slovníkové heslo (viz též obrázek 4.6) obsahuje vedle technických údajů tyto informace:

BASIC_FORM je záhlaví hesla, obsahuje základní tvar víceslovné jednotky. Některá slova výrazu mohou být shodná se svými lemmaty (tj. infinitiv, nominativ, singulár, …), ale nemusí to platit pro všechny části VV, například slovesný VV ve větě *„Svým ohnivým projevem strhl publikum na svou stranu.“* má svou BASIC_FORM *„strhnout na svou stranu“*, kde substantivum je v akuzativu.

LEMMATIZED je základní tvar hesla (BASIC_FORM) lemmatizovaný po jednotlivých slovech. Předchozímu příkladu odpovídá *„strhnout na svůj strana“*. Údaj může posloužit pro identifikaci VV v lemmatizovaném textu.

MORPHO_TAGS Třetině VV byla v průběhu přípravy slovníku přiřazena morfologická informace pro každé jednotlivé slovo, zapsaná jednoduše jako řada morfologických značek v Pražském pozičním tagsetu.[55] Z kombinace LEMMATIZED a MORPHO_TAGS je tudíž snadné vygenerovat BASIC_FORM.

↪ *Příklad:* Pro VV *„strhnout na svou stranu“* vypadá řada značek následovně:

`Vf--------A---- RR--4---------- P8FS4---------1 NNFS4-----A----`

Morfologické značky však byly získány automatickou anotací, informace tudíž není natolik spolehlivá, aby byla systematicky uváděna u všech položek slovníku.

POS Kategorie slovního druhu je přiřazena víceslovnému výrazu jako celku. VV vystupují jako jednotky jazykového popisu, často je můžeme nahradit jediným slovem – a slovní druh popisuje funkci celého VV ve větě. Nijak nesouvisí (zejména u adverbií) se slovními druhy jeho částí.

↪ *Příklad: stůj co stůj*, *na zelené louce* i *jako když střelí do hejna vrabců* jsou adverbia.

GLOSS Pokud to posloužilo snazší srozumitelnosti slovníkového hesla, měli anotátoři možnost vyplnit glosu. Není to však položka povinná a u 96 % hesel chybí.

↪ *Příklad:* BASIC_FORM: *liturgický jazyk*, GLOSS: *jazyk používaný při bohoslužbě*

EXAMPLE Podobně mohli vyplnit také příklad užití, typicky tedy okopírovat větu, ve které se VV vyskytuje a kvůli níž zakládají nové slovníkové heslo. (Pouze tři procenta VV obsahují příklad.)[56]

SYNONYMS Některá hesla (přibližně 9 %, obvykle ta, která pochází z Českého WordNetu či z EuroVocu), obsahují jeden či více synonymních výrazů. Tyto výrazy mohou, ale nemusí být víceslovné.

↪ *Příklad:* VV *studijní program* obsahuje synonyma *osnova*, *učební plán* a *učivo*.

PDT25_FREQ V této položce je uložen počet výskytů daného VV v datech korpusu PDT 2.0. Jak bylo řečeno, slovník je sestaven právě jen z VV vyskytujících se v PDT, tudíž všechny hodnoty jsou nenulové.

[55] `https://ucnk.ff.cuni.cz/doc/popis_znacek.pdf`

[56] Někdy je věta vyplněna, pokud se jedná o nečekaně kreativní užití VV, jako například *„Američtí sloni v ulsterském porcelánu“*.

SOURCE je jeden z technických, méně důležitých údajů a říká, z kterého slovníku či od kterého anotátora daná položka pochází. Dále je ve slovníku čas vzniku, informace o změně položky atp.

TREE_STRUCTURE je položka, ve které se skrývá největší přidaná hodnota SemLexu. Obsahuje informaci o vnitřní syntaktické struktuře víceslovného výrazu. V základní, zveřejněné verzi slovníku obsahuje informace z t-roviny PDT. Jsou to všechny t-uzly reprezentující VV: jejich `t_lemma`ta a jejich vzájemné závislostní vztahy (viz též str. 141). Díky propojení slovníku s daty jsme však vytvořili pro potřeby experimentu popisovaného v sekci 6.1.6 verzi slovníku se stromovou strukturou pocházející z a-roviny. Podobně pokud se to ukáže jako potřebné, lze v dalších verzích slovníku doplnit do snadno rozšiřitelného záznamu například informaci o funktorech, nebo informaci z jiné roviny jazykového popisu (kupř. pomocné uzly z a-roviny, tzv. odkazy `aux`).

Hesla SemLexu nejsou vzájemně propojována tak, jako je tomu v lexikálních databázích typu WordNet či FrameNet, nicméně tři druhy odkazů zde přeci jen existují. Prvním druhem je odkaz na synonymní slovníkové heslo, druhým je heslo odvozené z jiného pomocí jazykové analogie, často jednorázově vymyšlené právě pro ten případ[57] a třetím je obecný nespecifikovaný odkaz na související, příbuzný VV. Takto například vede odkaz od *pohlavně zneužívat* k heslům *pohlavně zneužít* (tedy vidový protějšek víceslovného výrazu) a *pohlavní zneužívání* (tedy nominalizovaný VV). Tento třetí typ odkazu jsme vybrali též do ukázky na obrázku 4.6.

4.9.1 Formát SemLexu

Celý SemLex o 8 797 VV má velice jednoduchou strukturu: pole 8 797 prvků, kde každý prvek tvoří asociativní pole atributů a jejich hodnot. Takto prostou datovou strukturu lze serializovat různým způsobem, SemLex k tomu využívá formátu YAML,[58] který umožňuje přímo reprezentovat datové struktury konkrétního programovacího jazyka, v našem případě Perlu. Uložení i opětovné načtení celého slovníku je tudíž snadné, neboť se ukládá přímo datová struktura.

V následující ukázce jediného hesla (*vojsko rychlého nasazení*, které jsme už viděli na obrázku 4.6) značí úvodní pomlčka, že se jedná o další položku pole, a zbytek řádku zaznamenává, že tato položka je asociativní pole (hash) perlové

[57] Například *akademická vesnice* ve větě *„Očekává se, že studentských her se zúčastní 5 100 sportovců a 100 funkcionářů ze 135 zemí, akademická vesnice však má jen 4 570 míst.“* anotátor označil jako „derivované z hesla *atletická vesnice*“.

[58] `http://yaml.org`

třídy `SemLex_heslo`. Následují dvojice atribut – hodnota, které přímo odpovídají údajům z obrázku 4.6 a které jsou (až na jedinou) srozumitelné bez dalších slov.

```
1  - !!perl/hash:SemLex_heslo
2    BASIC_FORM: vojsko rychlého nasazení
3    CREATED: 080816140714
4    EXAMPLE: ''
5    GLOSS: souvisí s 0000006200 (brigáda r.n.) a 0000032241 (síla r.n.)
6    ID: 0000012893
7    LEMMATIZED: vojsko rychlý nasazení
8    MODIFIED: 090607123501
9    MODIFIER: annotA_annotB_merge
10   MORPHO_TAGS: ''
11   ORIGID: ~
12   PDT25_FREQ: 1
13   POS: 'N'
14   SOURCE: annotA_annotB
15   SYNONYMS: ~
16   TREE_STRUCT:
17     -
18       - vojsko
19       - ~
20     -
21       - nasazení
22       - 0
23     -
24       - rychlý
25       - 1
```

Technické informace (například časová razítka) a položky, které pro toto heslo nejsou vyplněny, jsme zobrazili šedou barvou.

Jediná část, kterou je potřeba vysvětlit, je poslední položka TREE_STRUCTURE (od řádku 16). Celý strom je reprezentovaný polem uzlů: každý obsahuje své `t_lemma` a index uzlu, na němž je závislý. (Indexy jsou číslovány od 0, kořen má místo indexu otce symbol ~ značící `undef`.)

4.10 VALLEX ve formátu verze B

V následujících sekcích se vracíme k prvním dvěma slovníkům, VALLEXu a PDT-Vallexu, představeným na začátku kapitoly (sekce 4.2 a 4.3) a budeme se jim věnovat z hlediska jejich datových formátů.

VALLEX byl od počátku navrhován tak, aby ho mohl pohodlně používat jak člověk, tak i počítač. Proto byl uživatelům mimo webovou verzi s možností vyhledávání podle četných kritérií a mimo tisknutelnou verzi (PDF) již od verze 1.0 k dispozici také ve formátu XML. Právně na něj se zde soustředíme, neboť tato práce se zabývá automatickým zpracováním. Tento formát pak doznal zásadní obměny a spolu s verzí VALLEX 2.5 byl vydán ve formátu nazývaném „verze B“ (Žabokrtský, 2005).

Nejdůležitější odlišnosti oproti předchozí verzi spočívají v zacházení s lexémem, jehož přímý záznam v XML verze B mnohem více odpovídá teorii FGP. Podrobně vysvětlíme dvě změny, které mají vliv na strukturu slovníku a vedly nás k zavedení nového formátu SAMR a převedení dat do tohoto formátu, viz sekce 4.13:

1. Reflexivní částice se připisuje k lexému jako celku.
2. Popis vidových dvojic tvořících lexém je společný.

1. Reflexivní částice. Reflexivní částice je ve verzi B připsaná k celému lexému, nikoli zvlášť ke každému m-lemmatu, které sdružuje,[59] jak tomu bylo v předchozích verzích a jak tomu je u většiny jiných slovníků.

↪ *Příklad:* První část záznamu pro lexém *namáčet se*^{impf} *namočit se*^{pf} vypadá takto:

namáčet se, namočit se

```
<lexeme pos='v' id='lxm-v-namáčet-se-namočit-se'>
    <lexical_forms>
        <mlemma aspect='impf' coindex='impf'>namáčet</mlemma>
        <mlemma aspect='pf' coindex='pf'>namočit</mlemma>
        <reflex>se</reflex>
    </lexical_forms>
    ...
```

Ačkoli `mlemma` (morfologické lemma) je uvedeno jen jako *namáčet*, spolu s `reflex` *se* je ovšem chápáno jako reflexivní, tedy lemma *namáčet se*. (Totéž pro m-lemma *namočit*.)

Teoreticky přesný popis (reflexivní částici skutečně vždy sdílí všechna lemmata sdružená v lexém) je také kompaktnější a datová reprezentace je úspornější. Nevýhodou ovšem je, že se nikde nevyskytuje lemma dohromady se svou reflexivní částicí (tedy *„namočit se“*) a veškeré nástroje pracující s tímto XML si musí přítomnost částice zjistit a případně ji připojit k m-lemmatu, pokud to potřebují, ve své interní reprezentaci.

[59] Množinu lemmat v lexému popisuje bod dva.

2. Lexém sdružuje vidové protějšky. Byla změněna základní struktura uspořádání VALLEXu[60] a celý lexém byl vnořen do jednoho elementu. (Toto členění odráží přístup FGP k lexému jako k základní jednotce, kterou má smysl popisovat. Jejími reprezentanty jsou různé tvary slova, a to *včetně* různých vidových protějšků, viz Panevová, Benešová a Sgall, 1971.) Ve verzi B je tedy hlavním XML elementem `lexeme`, který sdružuje všechny ortografické varianty i vidové protějšky, zatímco v předchozí verzi byl element `word_entry` zvlášť pro každý vid a tato hesla mezi sebou byla jen provázána pomocí odkazů na vidový protějšek.

Tento přístup se odrazil i ve slovnících tištěném a webovém, kde je lexém od verze 2.5 také základním slovníkovým heslem. I do těchto podob slovníku se tedy promítl základní přístup FGP.

Více lexémů může být sdruženo do *lexémového shluku* (`lexeme_cluster`), který spojuje lexémy se stejným m-lemmatem lišící se pouze reflexivní částicí (např. *„mýlit (si/se)“*). Tento XML element již nemá žádný odraz v jiné než datové podobě slovníku.

Kvůli změně struktury – spojení vidových protějšků do společných lexémů – bylo nutné zavést dva nové mechanismy:

- *Ko-indexace* každého příkladu a každé glosy s tím konkrétním lemmatem, ke kterému v lexému patří. *Příklad:* Pro lexém *namáčet se*^{impf} *namočit se*^{pf} je u první lexikální jednotky uveden její význam v glose takto:

namáčet se, namočit se, LU č. 1

```
18      <gloss>
19          <coindexed coindex='impf'>máčet se</coindexed>
20          <coindexed coindex='pf'>zmáčet se</coindexed>
21      </gloss>
```

  Přepsáno pro lidského uživatele (tak, jak je to používáno v tištěném slovníku i na webu) to vypadá takto:

  glosa: *impf* máčet se; *pf* zmáčet se.

  Atribut `coindex` na řádcích 19 a 20 přiřazuje glosy k `mlemma`tům na řádcích 3 a 4 v předchozím příkladu. Glosa *„zmáčet se“* se tedy vztahuje k lexikální

[60] Zde VALLEXem rozumíme jeho datovou reprezentaci v XML formátu v souboru `vallex-2.5.xml`. Strukturou se tedy myslí jeho linearizace při zápisu, rozčlenění do hierarchicky uspořádaných XML elementů. To je samozřejmě izomorfní s jinými zápisy a na ně převoditelné. (Například izomorfní s formátem VALLEXu 1.0 – již tam byl lexém v popisované podobě dostupný skrze odkazy na vidové protějšky). Tady nicméně mluvíme o zvoleném zápisu, v němž je slovník šířen a jehož základem je lexém.

formě *„namočit se“*, jejíž vid (jak vidíme až na řádku 4) je perfektivní, dokonavý.[61]

Pokud lexém obsahuje ortografické varianty nějakého lemmatu, jsou ko-indexovány společně jako v případě lexému *chytat se*^{impf}*, chytit se/chytnout se*^{pf} na straně 64, řádek 4.

- *Omezení* jen na některé vidy. Ačkoli lexikální jednotky se prototypicky vztahují k celému lexému, tedy ke všem lexikálním formám v něm, někdy nastává situace, kdy LU existuje pouze pro některý vid,[62] nebo pro jedno z více lemmat vyjadřující stejný vid,[63] či dokonce pouze pro některou ortografickou variantu lemmatu.[64] V takovém případě se toto omezení LU uvádí výčtem povolených (vidových, nebo ortografických) variant. (Připomeňme, že skupiny ortografických variant jsou ko-indexovány jako celek, není tedy možné se na ně pouze odkázat.) *Příklad:* Pro jednu LU ukázkového lexému je mj. uvedeno:

namáčet se, namočit se, LU č. 2

```
29  <blu id='blu-v-namáčet-se-namočit-se-2'>
30      <lexical_forms>
31          <mlemma aspect='pf' coindex='pf'>namočit</mlemma>
32          <reflex>se</reflex>
33      </lexical_forms>
34      <gloss>
35          zaplést se do nepříjemností
36      </gloss>
37      <example>
38          namočil se do pěkného maléru
39      </example>
40      ...
```

[61] Samotná hodnota atributu `coindex` se od vidu může mírně lišit v případě, že do lexému patří více lemmat stejného vidu. *Příklad:* V lexému obsahujícím *oloupávat*^{impf}*, olupovat*^{impf}*, oloupnout*^{pf}*, oloupat*^{pf} musí být příklad spojený s konkrétní perfektivní lexikální formou, což se rozlišuje právě pomocí ko-indexu:

```
<example>
    <coindexed coindex='pf1'>oloupat brambory</coindexed>
    <coindexed coindex='pf2'>ukaž, oloupnu ti ten pomeranč</coindexed>
    ...
```

[62] Např. lexikální jednotka s významem „uvádět přehnané odhady“ v rámci lexému *házet*^{impf}*, hodit*^{pf} nejde použít jako **Včera jsi zas v hospodě hodil čísly*, ale výhradně *…házel čísly.*

[63] Např. pro lexém *zasunovat*^{impf}*, zasouvat*^{impf}*, zasunout*^{pf}*, zasout*^{pf} ve významu „překrývat“ lze říci *Tu vzpomínku jsem zasul hluboko*, ale nikoli **Tu vzpomínku jsem zasunul hluboko.*

[64] Např. pro *odepisovat/odpisovat*^{impf}*, odepsat*^{pf} ve významu „písemně odpovídat“ nelze použít *odpisovat.*

Tedy základní lexikální jednotka (`blu`) s číslem dva a významem *„zaplést se do nepříjemností“* (ř. 35) přepisuje řádkem 31 množinu lexikálních forem pouze na *namočit se*^{pf}.[65] Protože se nyní tato LU vztahuje jen k jediné lexikální formě, není již potřeba uvádět ko-index v glose a příkladu, jako tomu bylo na řádkách 19 a 20.

Sdružení vidových protějšků do jednoho lexému má teoretické opodstatnění, ale také pomáhá udržet konzistentní informace u vidových dvojic, které by se jinak editovaly odděleně a docházelo by k chybám.

Na závěr ještě připomeňme, že již od verze 1.0 jsou stejným způsobem sdružené ortografické varianty a od formátu verze B se sjednotilo jejich zachycení se zachycením vidových protějšků.

Jakkoli je formát VALLEXu verze B teoreticky výstižný a technicky propracovaný, s malými nároky na prostor, jeho velkou nevýhodou je komplikovaná práce s ním. Při každém zpracování či i jednoduchém dotazu je potřeba kontrolovat mnoho informací uložených na různých místech příslušného záznamu.

↪ *Příklad:* V případě ortografické varianty reflexivního slovesa *chytnout se* je atribut `aspect` i samotný element `mlemma` (řádek 6) na jiné úrovni hierarchické XML struktury než `mlemma` pro imperfektivní lemma (ř. 3), zatímco atribut `coindex` je na stejné úrovni, ale pokaždé u jiného elementu (3 a 4).

chytat se, chytit se, chytnout se

```
1  <lexeme pos='v' id='lxm-v-chytat-se-chytit-se-chytnout-se'>
2      <lexical_forms>
3          <mlemma aspect='impf' coindex='impf'>chytat</mlemma>
4          <mlemma_variants coindex='pf'>
5              <mlemma aspect='pf'>chytit</mlemma>
6              <mlemma aspect='pf'>chytnout</mlemma>
7          </mlemma_variants>
8          <reflex>se</reflex>
9      </lexical_forms>
10     ...
```

Chceme-li všechna tři slovesa zpracovávat společně a navíc jako reflexivní (řádek 8), musíme tak ošetřit mnoho speciálních případů.

[65] Opět může nastat případ, že z více zástupců jednoho vidu bude lexikální jednotka platná pouze pro jeden, určený ko-indexem. *Příklad: spadat-I*^{impf}*, spadávat*^{impf}*, spadnout*^{pf}*, spadat-II*^{pf}: Kurz měny může *spadnout*, ale nemůže **spadat*; naopak listí ze stromu na podzim nevídáme **spadnout*, ale právě *spadat* – ve všech případech jde o dokonavý vid. Tudíž je nutné rozlišit (pomocí ko-indexu), na které lemma se daná LU vztahuje.

Na příkladu sloves *svírat, sevřít* (obrázek 4.1 na straně 26) jsme navíc mohli vidět, že ačkoli byly informace v jednotlivých LU pro oba vidové protějšky téměř shodné, s výjimkou valenčních rámců a sémantické třídy byly (příklady, glosy apod.) stejně vždy uvedeny odděleně, neboť se lišily právě videm.

Doplňme, že existují i řídké případy, kdy je sloučení lemmat do lexému vysloveně kontraproduktivním. Takovým příkladem jsou slovesa *propadávat*^{impf}, *propadat-I*^{impf}, *propadnout*^{pf}, *propadat-II*^{pf}. Ze všech devíti lexikálních jednotek se lemmatu propadat-*II*^{pf} týká jen jediná (~ postupně propadat: „všechen popel propadal“) a ta naopak není použitelná pro žádné jiné z lemmat. Proč tedy toto lemma slučovat s ostatními?

Tyto důvody nás vedly k vytvoření zjednodušeného (ale izomorfního) formátu SAMR, který popíšeme v sekci 4.13.

4.11 Formát PDT-Vallexu

Pro anotaci PDT 2.0, během níž vznikal PDT-Vallex, byl vyvinut editor TrEd (Pajas a Štěpánek, 2008)[66] spolu s dotazovacím jazykem PML-TQ (Pajas a Štěpánek, 2009)[67] a také nový datový formát PML.

Prague Markup Language (Pajas a Štěpánek, 2005)[68] je formát s vlastním schematem založený na XML. PML schema přináší novou úroveň abstrakce nad standardními jazyky pro popis XML schemat jako je DTD, RelaxNG, nebo W3C XML Schema, a tím také přesnější specifikaci než tato schemata, viz Pajas a Štěpánek (2005, str. 15). Umožňuje totiž přisoudit jednotlivým elementům tzv. „role“, které jsou ortogonální k datovým typům, jako je seznam, výčet, alternativy apod. Takové role slouží jako vodítko pro aplikace, které data zpracovávají. Příkladem rolí je třeba `#NODE` pro XML elementy reprezentující uzly stromů, `#ORDER` pro elementy určující pořadí uzlů (které tak nemusí souviset s uspořádáním uzlů v souboru), `#HIDE` pro elementy, které se nemají zobrazovat, nebo `#KNIT` pro určení míst, kam se mají na základě odkazů vložit části jiných XML dokumentů.

Ačkoli první použití PML bylo pro PDT,[69] umožňuje vytvořit schémata také pro slovníky. Ty je potom možné zobrazovat i editovat přímo ve výše zmíně-

[66] `http://ufal.mff.cuni.cz/tred/`

[67] `http://ufal.mff.cuni.cz/pmltq/`

[68] `http://ufal.mff.cuni.cz/jazz/PML/`

[69] PDT je rozdělené do jednotlivých rovin jazykového popisu a jedna z věcí, kterou PML nabízí, je rozdělení do více souborů provázaných odkazy, které lze potom při práci se soubory např. v TrEdu tzv. „sešít“ (právě pomocí role `#KNIT`) a používat jako celek. Hierarchická stromová struktura analytické a tektogramatické roviny je v PML zachycena přirozeným prostředkem hierarchického XML stromu.

ném TrEdu. Proto i PDT-Vallex byl záhy po začátku anotací uložen ve formátu PML.[70]

Pokud odhlédneme od výhod plynoucích z PML, můžeme PDT-Vallex zpracovávat také jako každé jiné XML. Výpis formátu uvádíme v ukázce na str. 69. Tělo dokumentu obsahuje elementy `word`, tedy slovníková hesla, která, jak jsme viděli, v PDT-Vallexu odpovídají lemmatům. Ty neobsahují nic nečekaného ani nic, co by nebylo na první pohled srozumitelné, snad jedině s výjimkou popisu morfo-syntaktických informací pro valenční doplnění, tedy obsah elementu `word/valency_frames/element/form`. V případě, že může být valenční doplnění tvořeno více slovy, je v elementu `form` několik v sobě zanořených elementů `node`, jejichž XML struktura reprezentuje závislostní syntaktickou strukturu, kterou musí dané uzly splňovat na analytické rovině. (Element `node` může být nahrazen v případě, že se jedná o stavové (prázdný uzel `state`) nebo typické (`typical`) doplnění pro daný funktor.) Každé další možné morfo-syntaktické doplnění je zapsáno jako další `form`. Můžeme tedy uvést ukázku zápisu obligatorního PATientu, která je srozumitelná bez dalšího vysvětlování, jedná se o zápis forem pro doplnění PAT(o+4,inf,aby), kde vedlejší věta uvozená spojkou *aby* je znázorněna závislým slovesem:

```
<element functor="PAT" type="oblig">
<form>
  <node afun="unspecified" agreement="0" lemma="o-1">
    <node afun="unspecified" agreement="0" case="4"/>
  </node>
</form>
<form>
  <node afun="unspecified" agreement="0" pos="f"/>
</form>
<form>
  <node afun="unspecified" agreement="0" lemma="aby">
    <node afun="unspecified" agreement="0" pos="v"/>
  </node>
</form>
</element>
```

[70] VALLEX je také převoditelný do PML formátu. V tomto formátu ho lze otevřít v TrEdu a využít PML-TQ pro vyhledávání složitých mnohakriteriálních dotazů, viz Bejček, Kettnerová a Lopatková (2010).

Podali jsme základní vysvětlení zápisu morfo-syntaktické informace a celého formátu PDT-Vallexu podle PML schematu `vallex_schema.xml`. Další vlastnosti vysvětlíme vždy na příslušných místech, kde je budeme potřebovat.

4.12 Porovnání formátů VALLEXu a PDT-Vallexu

Viděli jsme, že formát VALLEXu je poměrně složitý (4.10) a nabízela se otázka, zda by se nevyplatilo převést ho do formátu jednoduššího.

Druhým důvodem pro převod je skutečnost, že formáty slovníku VALLEX a PDT-Vallex jsou možná až zbytečně odlišné, přihlédneme-li k tomu, že popisují stejné fenomény na základě stejné teorie. Již jen proto je tedy výhodné hledat formát společný. Než tedy (v následující sekci 4.13) přistoupíme k samotnému popisu sjednocujícího formátu SAMR, srovnejme ukázku jednoho lexému (obsahujícího dvě lemmata *„sevřít, svírat“*) v obou slovnících.[71] Jsou to tatáž slovesa, která jsme ukazovali v předchozích sekcích na obrázcích 4.1 (VALLEX, str. 26) a 4.2 (PDT-Vallex, str. 32) v čitelnější podobě z webového rozhraní.

První ukázka je z VALLEXu. Popisuje tři lexikální jednotky (řádky v9, v44 a v64) společné pro celý lexém, ačkoli u druhé LU je upřesněno, že se vztahuje pouze na lemma *sevřít* (v46). Kromě valenčního rámce (v10 až v23) popisuje první LU ještě reflexivitu (od v32),[72] reciprocitu (od v36) a sémantickou třídu (v40), druhá umožňuje pouze reflexivitu (v60) a třetí lexikální jednotka je idiomatická (v63).

Celý lexém je pak ještě společně s reflexivním lexémem *„sevřít se, svírat se“* (v88) uzavřen do lexémového shluku (v1).

Ukázka sloves sevřít, svírat *ve VALLEXu*

```
<lexeme_cluster>
  <lexeme pos='v' id='lxm-v-sevřít-svírat'>
    <lexical_forms>
      <mlemma aspect='impf' coindex='impf'>svírat</mlemma>
v5    <mlemma aspect='pf' coindex='pf'>sevřít</mlemma>
    </lexical_forms>
    <lexical_units>
      <lu_cluster id="luc-v-sevřít-svírat-1">
        <blu id='blu-v-sevřít-svírat-1'>
v10       <frame>
            <slot functor='ACT' type='obl'>
              <form type="direct_case" case="1" />
            </slot>
```

[71] Části, které nepovažujeme za důležité, jsou vytištěny šedě, aby byl zbytek kódu o trochu přehlednější.

[72] Od verze 2.6.3 jsou tyto elementy přejmenovány na `<diat type="deagent">` v rámci systematického zpracování diatezí a lexikalizovaných alternací.

```
            <slot functor='PAT' type='obl'>
v15           <form type="direct_case" case="4" />
            </slot>
            <slot functor='EFF' type='opt'>
              <form type="prepos_case" prepos_lemma="do" case="2" />
              <form type="prepos_case" prepos_lemma="v" case="4" />
v20         </slot>
            <slot functor='LOC' type='typ' />
            <slot functor='MANN' type='typ' />
          </frame>
          <gloss>
v25         <coindexed coindex='impf'>těsně spojovat; tisknout; objímat</coindexed>
            <coindexed coindex='pf'>těsně spojit; stisknout; obejmout</coindexed>
          </gloss>
          <example>
            <coindexed coindex='impf'>svíral ruku v pěst; pravá ruka svírala hůl</coindexed>
v30         <coindexed coindex='pf'>sevřel ruku v pěst; sevřel ji v náručí</coindexed>
          </example>
          <rfl type="pass">
            <coindexed coindex='impf'>ruka se svírala v pěst</coindexed>
            <coindexed coindex='pf'>ruka se sevře v pěst</coindexed>
v35       </rfl>
          <rcp type="ACT-PAT">
            <coindexed coindex='impf'>svírali se v náručí</coindexed>
            <coindexed coindex='pf'>sevřeli se v náručí</coindexed>
          </rcp>
v40       <class>change</class>
        </blu>
      </lu_cluster>
      <lu_cluster id="luc-v-sevřít-svírat-2">
        <blu id='blu-v-sevřít-svírat-2'>
v45       <lexical_forms>
            <mlemma aspect='pf' coindex='pf'>sevřít</mlemma>
          </lexical_forms>
          <frame>
            <slot functor='ACT' type='obl'>
v50           <form type="direct_case" case="1" />
            </slot>
            <slot functor='PAT' type='obl'>
              <form type="direct_case" case="4" />
            </slot>
v55         <slot functor='LOC' type='typ' />
          </frame>
          <gloss>chytit; obklíčit</gloss>
          <example>policie sevřela demonstranty na Národní třídě;
            irácké jednotky sevřeli spojenci</example>
v60       <rfl type="pass">demonstranti se sevřou na Národní třídě</rfl>
        </blu>
      </lu_cluster>
      <lu_cluster id="luc-v-sevřít-svírat-3" idiom="1">
        <blu id='blu-v-sevřít-svírat-3'>
v65       <frame>
            <slot functor='ACT' type='obl'>
              <form type="direct_case" case="1" />
            </slot>
            <slot functor='PAT' type='obl'>
v70           <form type="direct_case" case="4" />
            </slot>
            <slot functor='BEN' type='typ'>
              <form type="direct_case" case="3" />
```

```
            </slot>
v75       </frame>
          <gloss>
            <coindexed coindex='impf'>zachvacovat; působit nepříjemný pocit</coindexed>
            <coindexed coindex='pf'>zachvátit; způsobit nepříjemný pocit</coindexed>
          </gloss>
v80       <example>
            <coindexed coindex='impf'>hrůza jí svírala srdce</coindexed>
            <coindexed coindex='pf'>hrůza jí sevřela srdce</coindexed>
          </example>
        </blu>
v85   </lu_cluster>
    </lexical_units>
  </lexeme>
  <lexeme pos='v' id='lxm-v-sevřít-se-svírat-se'>
    ...
v90 </lexeme>
</lexeme_cluster>
```

Konec VALLEXu

Druhá ukázka obsahuje tatáž lemmata, tentokrát ovšem odděleně, neboť PDT-Vallex vidové protějšky nijak nespojuje. První je sloveso *sevřít* (od řádku p1) s jedinou lexikální jednotkou (začínající na p3. řádku). Sloveso *svírat* (p23) má lexikální jednotky tři (p25, p43 a p61; čtvrtá LU na řádku p81 je již neplatná, neboť byla nahrazena (`substituted`) třetí LU). Druhá LU pak odpovídá jediné LU prvního slovesa *sevřít*: *„svírat syna v náručí“* (p45) a *„sevřít syna v náručí“* (p5); tato korespondence ovšem ve slovníku není nikterak vyznačena.

Všechny lexikální jednotky obsahují příklad, poznámku s glosou a samotný valenční rámec, který je rozepsaný po jednotlivých uzlech syntaktického stromu.

Ukázka sloves `sevřít, svírat` *v PDT-Vallexu*

```
<word POS="V" lemma="sevřít" id="v-w6009">
  <valency_frames>
    <frame hereditary_used="2" status="reviewed" used="2" id="v-w6009f1"
        pdt_used="2" pdt_hereditary_used="2" pcedt_used="1" pcedt_hereditary_used="1">
 p5   <example>sevřít syna v náručí</example>
      <note>obejmout</note>
      <frame_elements>
        <element functor="ACT" type="oblig">
          <form>
p10         <node afun="unspecified" agreement="0" case="1" inherits="1" neg="unspecified"/>
          </form>
        </element>
        <element functor="PAT" type="oblig">
          <form>
p15         <node afun="unspecified" agreement="0" case="4" inherits="1" neg="unspecified"/>
          </form>
        </element>
      </frame_elements>
      <local_history/>
p20   </frame>
  </valency_frames>
</word>
<word POS="V" lemma="svírat" id="v-w6667">
  <valency_frames>
p25   <frame hereditary_used="1" status="reviewed" used="1" id="v-w6667f2"
```

```
      pdt_used="1" pdt_hereditary_used="1" pcedt_used="0" pcedt_hereditary_used="0">
    <example>přímky svírají úhel</example>
    <note>vymezovat, vytínat</note>
    <frame_elements>
      <element functor="ACT" type="oblig">
        <form>
          <node afun="unspecified" agreement="0" case="1" inherits="1" neg="unspecified"/>
        </form>
      </element>
      <element functor="PAT" type="oblig">
        <form>
          <node afun="unspecified" agreement="0" case="4" inherits="1" neg="unspecified"/>
        </form>
      </element>
    </frame_elements>
    <local_history/>
  </frame>
  <frame hereditary_used="1" status="reviewed" used="1" id="v-w6667f1"
      pdt_used="1" pdt_hereditary_used="1" pcedt_used="0" pcedt_hereditary_used="0">
    <example>svírat syna v náručí</example>
    <note>objímat</note>
    <frame_elements>
      <element functor="ACT" type="oblig">
        <form>
          <node afun="unspecified" agreement="0" case="1" inherits="1" neg="unspecified"/>
        </form>
      </element>
      <element functor="PAT" type="oblig">
        <form>
          <node afun="unspecified" agreement="0" case="4" inherits="1" neg="unspecified"/>
        </form>
      </element>
    </frame_elements>
    <local_history/>
  </frame>
  <frame id="v-w6667f3_ZU" status="active" pdt_used="0" pdt_hereditary_used="0"
      pcedt_used="1" pcedt_hereditary_used="1">
    <example>Rozpočtovými tlaky svíraly společné projekty stále více.</example>
    <note>omezovat, poškozovat</note>
    <frame_elements>
      <element functor="ACT" type="oblig">
        <form>
          <node case="1" inherits="1"/>
        </form>
      </element>
      <element functor="PAT" type="oblig">
        <form>
          <node case="4" inherits="1"/>
        </form>
      </element>
    </frame_elements>
    <local_history>
      <local_event time_stamp="12.9.2011 18:10:09" type_of_event="create" author="ZU"/>
    </local_history>
  </frame>
  <frame status="substituted" id="v-w6667hsa_431" substituted_with="v-w6667f3_ZU"
      pdt_used="0" pdt_hereditary_used="0" pcedt_used="0" pcedt_hereditary_used="0">
    <example>Společné projekty jsou stále více svírány rozpočtovými tlaky.</example>
    <frame_elements>
      <element functor="ACT" type="oblig">
```

```
      ...
    </frame>
  </valency_frames>
</word>
```

Konec PDT-Vallexu

Oba slovníky tedy skutečně popisují téměř stejnou informaci podobným způsobem. Hlavní rozdíly si popíšeme rozdělené na věcné a technické rozdíly.

1. Věcné rozdíly. Věcnými rozdíly myslíme odlišnou informaci, kterou slovníky zachycují – tedy takové části slovníku, které na sebe nejsou žádným způsobem převoditelné. Jsou to

(i.) delimitace významů do LU,
(ii.) informace o morfo-syntaktické struktuře valenčních doplnění,
(iii.) počet instancí LU v datech,
(iv.) informace o reciprocitě, reflexivitě, (idiomaticitě).

(i.) Každý slovník vymezuje **významy** (a odlišné valenční rámce) jinak. Ve VALLEXu jsou v jednom lexému popsány tyto lexikální jednotky:

$sevřít_1^{pf\ V}$	~ stisknout, obejmout	… *sevřel ruku v pěst / dceru v náručí*
$sevřít_2^{pf\ V}$	~ obklíčit	… *policie sevřela demonstranty*
$sevřít_3^{pf\ V}$	~ zachvátit	… *hrůza jí sevřela srdce*
$svírat_1^{impf\ V}$	~ tisknout, objímat	… *svíral ruku v pěst / dceru v náručí*
$svírat_3^{impf\ V}$	~ zachvacovat	… *hrůza jí svírala srdce*

a v PDT-Vallexu jsou pro stejnou dvojici lemmat vymezeny následující „valenční rámce“ (ve smyslu významu), tedy lexikální jednotky v terminologii VALLEXu:

$sevřít_1^{pf\ P}$	~ obejmout	… *sevřela syna v náručí*
$svírat_2^{impf\ P}$	~ vytínat	… *přímky svírají úhel*
$svírat_1^{impf\ P}$	~ objímat	… *svírala syna v náručí*
$svírat_{3\_ZU}^{impf\ P}$	~ omezovat	… *svírali je finančními požadavky*

Vidíme dosti odlišnou delimitaci významů. Společný oběma slovníkům je význam *obejmout, objímat*, ovšem ve VALLEXu je spojen i s obecným tisknutím, mačkáním. Naopak vyděluje význam *obklíčit*, který v PDT-Vallexu nejspíš chybí (nebo by se musel vejít pod LU č. 1). Zbylé LU jsou zřejmě přítomny výlučně v jednom ze slovníků: *svírání hrůzou*, *svírání přímek* ani *svírání požadavky* nenacházíme v druhém slovníku.

(ii.) PDT-Vallex je mnohem podrobnější v zachycení povrchových vlastností valenčního rámce, neboť zachycuje také **syntaktické vztahy** mezi částmi jednoho valenčního doplnění. VALLEX zachycuje morfematické formy jako výčet (jako třeba pro EFF na řádcích v18 a v19 na straně 68). Kdyby v ukázce byl předložkový pád i pro PDT-Vallex, byly by XML elementy zanořeny ve shodě se syntaktickou závislostí uzlů, tedy předložka a na ní závisející jméno v daném pádě:

EFF(do+2,v+4) ve formátu PDT-Vallexu

```
<element functor="EFF" type="non-oblig">
  <form>
    <node agreement="0" lemma="do-1">
      <node agreement="0" case="2"/>
    </node>
    <node agreement="0" lemma="v-1">
      <node agreement="0" case="4"/>
    </node>
  </form>
</element>
```

To je informace, kterou z VALLEXu nelze vytěžit, zejména u složitějších případů, jako jsou třeba závislé části frazému, DPHR:[73]

na tenký led ve formátu PDT-Vallexu

```
<element functor="DPHR" type="oblig">
  <form>
    <node agreement="0" lemma="na-1">
      <node agreement="0" case="4" lemma="led" num="S">
        <node agreement="1" lemma="tenký"/>
      </node>
    </node>
  </form>
</element>
```

Jde o víceslovný výraz *„dostat se na tenký led“*. Zde je zachycen pád i číslo slova *led* a shoda se slovem *tenký*. Ve VALLEXu se v takovém případě uvádí pouze řetězec ”na tenký led”, tedy

```
<slot functor="DPHR" type="obl">
  <form type="phraseme_part" phraseme_part="na tenký led"/>
</slot>
```

(iii.) Protože je PDT-Vallex propojen s daty PDT 2.0, uvádí pro každou lexikální jednotku počet jejích **výskytů** v PDT a od rozšíření v PDT-Vallexu 2.0 také počet výskytů v české části korpusu PCEDT[74] (např. p25 na straně 69: 1× v PDT).

[73] V obou příkladech byly nepodstatné atributy vypuštěny.

[74] http://ufal.mff.cuni.cz/pcedt2.0/

(iv.) Naopak ve VALLEXu jsou doplňující informace o **reflexivitě** (např. v32 na straně 68), **reciprocitě** (v36), nebo třeba sémantické třídě lexikální jednotky.

2. Technické rozdíly. Technickými rozdíly myslíme takové části datové reprezentace slovníků, které zachycují stejnou informaci, nebo alespoň stejný typ informace, jen jsou v obou slovnících reprezentovány odlišně – nicméně vzájemně převoditelně. Technické rozdíly jsou:

(i.) logická struktura slovníkového hesla, reflexiva, ko-indexace,
(ii.) zápis valenčního rámce,
(iii.) pojmenování XML uzlů.

(i.) V první řadě mezi technickými rozdíly rozhodně stojí vlastnosti VALLEXu ve formátu verze B, o nichž jsme psali v sekci 4.10. Jedná se o celé záhlaví **slovníkového hesla**, které je v případě VALLEXu tvořeno lexémem a výčtem do něj náležejících m-lemmat (v3 na straně 67). Následuje případná reflexivní částice pro primární reflexiva tantum a odvozená reflexiva. Takto jsou všechny vidové protějšky pospolu. Je-li potřeba dále v lexikálních jednotkách uvést informaci specifickou jen pro některé z lemmat, odkáže se na ně ko-indexací, případně se zúží výčtem platných lemmat. Naproti tomu v PDT-Vallexu jsou jednotlivá lemmata striktně oddělena a nevede mezi nimi ani žádné jiné pojítko.

(ii.) PDT-Vallex má bohatší syntaktickou informaci o **valenčním rámci**, nicméně tato informace je v řadě případů (alespoň jednosměrně) převoditelná. Zápis je odlišný, ale můžeme jej transformovat (s využitím dodatečných informacích) do zápisu VALLEXu, jak je vidět níže. Vlevo je ukázka (ze slovesa *brát*) doplnění funktoru EFF. Méně podstatné informace jsou opět šedivé. Ty, které VALLEX vůbec nezachycuje, jsou zvýrazněny kurzivou: spojka *„jakožto“* se ve VALLEXu vůbec nevyskytuje, všude je pouze *„jako“* (4, 9); VALLEX také nijak nerozlišuje, zda za předložkami/spojkami stojí adjektivum, nebo nikoli (8, 19). Vpravo je tedy tato ukázka upravená tak, že jsou zmíněné doplňující informace specifické pro PDT-Vallex vypuštěny. Je to jakýsi hodně ochuzený formát PDT-Vallexu, mezikrok k převodu do formátu VALLEXu.

EFF(jako+4,za+4) v PDT-Vallexu

```
<element functor="EFF" type="oblig">
  <form>
    <node case="4" inherits="1">
      <node afun="AuxY" lemma="{jako,jakožto}"/>
    </node>
  </form>
  <form>
    <node case="4" inherits="1" pos="a">
      <node afun="AuxY" lemma="{jako,jakožto}"/>
    </node>
  </form>
  <form>
    <node lemma="za-1">
      <node case="4" inherits="1"/>
    </node>
  </form>
  <form>
    <node lemma="za-1">
      <node case="4" inherits="1" pos="a"/>
    </node>
  </form>
</element>
```

…bez specifických informací

```
<element functor="EFF" type="oblig">
  <form>
    <node case="4">
      <node lemma="jako"/>
    </node>
  </form>
  <form>
    <node case="4">
      <node lemma="jako"/>
    </node>
  </form>
  <form>
    <node lemma="za">
      <node case="4"/>
    </node>
  </form>
  <form>
    <node lemma="za">
      <node case="4"/>
    </node>
  </form>
</element>
```

Po takovém zjednodušení již jsou první dvě formy (2 a 7) stejné a druhé dvě (12 a 17) taktéž. Zbyly tedy již jen dvě různé formy; ty můžeme převést do zápisu VALLEXu. Musíme jen vědět, že dva vzájemně závislé uzly (předložka a na ní závislý uzel s určeným pádem; nebo uzel s určeným pádem a na něm závislá spojka *jako*) se oba zapisují shodně jako předložkové pády:

```
<slot functor="EFF" type="obl">
  <form type="prepos_case" prepos_lemma="jako" case="4"/>
  <form type="prepos_case" prepos_lemma="za" case="4"/>
</slot>
```

(iii.) Poslední drobný technický rozdíl je čistě v **pojmenování** XML uzlů (elementů a atributů). Například atribut `opt` vs. `non-oblig`; element `mlemma` asi zhruba odpovídá elementu `word`; element `gloss` se v PDT-Vallexu nazývá `note`.

Ukázali jsme na příkladu podobné i odlišné rysy formátů, v nichž jsou k dispozici slovníky VALLEX a PDT-Vallex. Také jsme naznačili možnosti transformace těchto formátů. Nyní popíšeme společný formát, který zachycuje oba slovníky.

4.13 Formát SAMR pro valenční slovníky

Z důvodů popsaných v předchozích sekcích jsme se rozhodli oba slovníky, VALLEX i PDT-Vallex, převést do společného formátu.

Za nejsnazší jsme uznali vytvoření vlastního formátu, co nejpodobnějšího oběma sjednocovaným XML formátům (pro zdůvodnění viz sekci 4.14). Navrhli jsme tedy XML formát, který vychází z formátu PDT-Vallexu a

- rozšiřuje ho o informaci o reflexivitě, reciprocitě, idiomaticitě a sémantické třídě,
- přidává možnost odkazů mezi jednotlivými slovníkovými hesly i mezi lexikálními jednotkami
- umožňuje variantu zápisu valenčního doplnění způsobem, který využívá VALLEX, neboť na závislostní zápis není převoditelný a
- přejmenovává některé XML elementy a atributy.

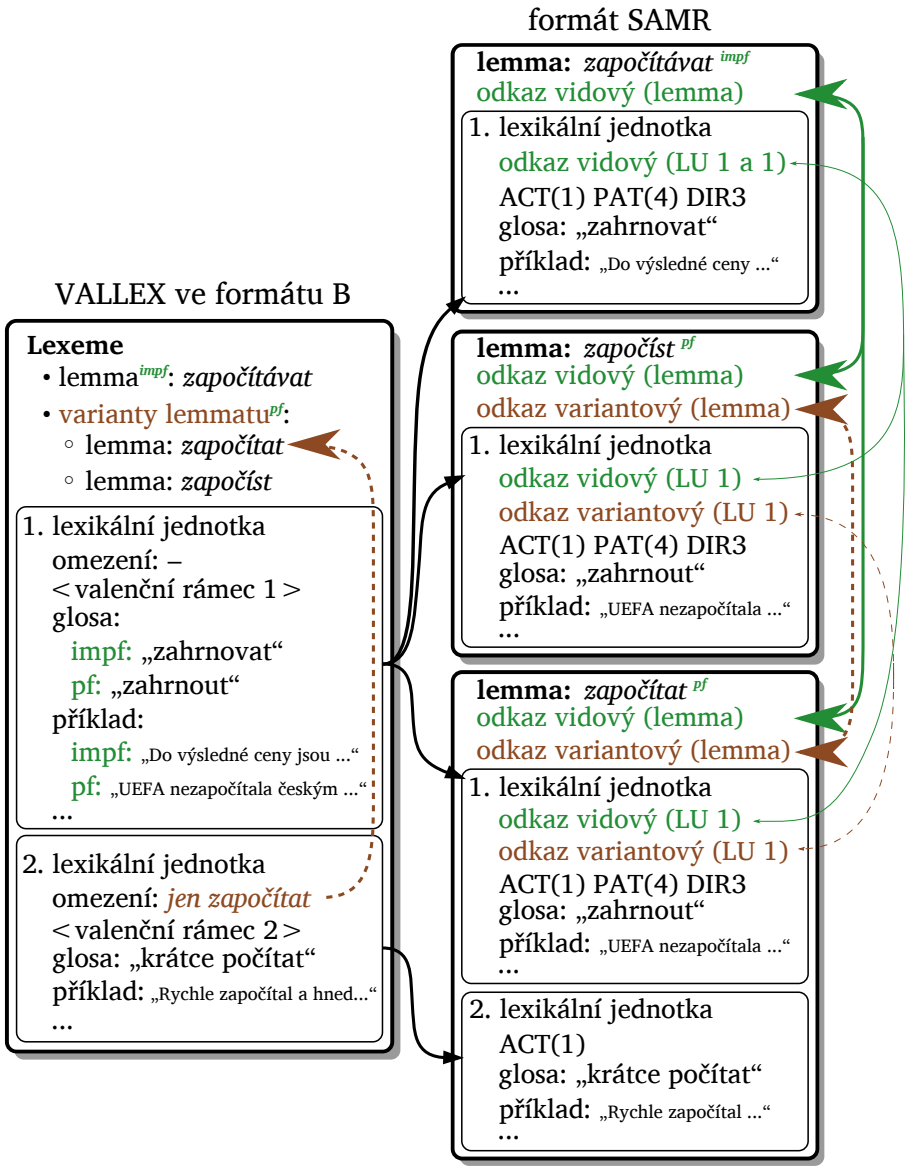
Převod PDT-Vallexu byl tedy přímočarý, spočíval jen v přejmenování elementů (a pro naši práci se slovesnou valencí také v odstranění všech slovníkových hesel pro substantiva, adjektiva a adverbia). Pro VALLEX (viz schematický obrázek 4.7) to znamenalo výrazný zásah, neboť bylo nutné *rozgenerovat* lexémy z formátu verze B do několika slovníkových hesel (jedno pro každý vidový protějšek, případně pro každou ortografickou variantu vidového protějšku). Tato slovníková hesla ovšem musí být mezi sebou provázána odkazy, aby byla zachována veškerá informace. Stejně tak byly rozgenerovány lexikální jednotky: z jedné LU náležející lexému vznikla LU náležející každému z hesel, která nahradila původní lexém, pokud pro dané heslo byla LU platná. Každé LU byla dále přiřazena patřičná ko-indexovaná část příkladu, glosy, příkladu na reciprocitu, apod. Také lexikální jednotky bylo nutné mezi sebou provázat pomocí odkazů, abychom zachovali korespondenci obdobných LU mezi vidovými protějšky. Informace vlastní pouze VALLEXu byly vloženy do nově vytvořených XML elementů (reflexivita, reciprocita, atd.) Tento formát jsme nazvali SAMR.[75]

Takto rozgenerovaná, *expandovaná* verze VALLEXu je z jednoho pohledu přechodem k formátu PDT-Vallexu, ovšem s mnohou doplněnou informací a bez syntaktických údajů u jednotlivých valenčních doplnění.

Zdůrazněme, že ačkoli hovoříme o jednotném formátu, v zápisu valenčních doplnění umožňuje obě možnosti: jak zápis nativní pro PDT-Vallex, tak i jednodušší pro VALLEX. V tomto místě tedy formát nic nesjednocuje a ani to nebylo jeho cílem: buď bychom informaci ztratili, nebo bychom ji museli na druhé straně doplnit ručně.

Zatímco složitost formátu VALLEXu verze B (tedy lexémy) jsme zredukovali za použití ekvivalentní úpravy formátu, složitost formátu PDT-Vallexu (tedy syntaktické informace) jsme zachovali. Ke ztrátě informace tedy nedochází.

[75] Zkratka ze Single Aspect, Multiple References. Vyjadřuje tedy stav, kdy každé slovníkové heslo již obsahuje jen jediný slovesný vid, zato může obsahovat celou řadu odkazů (na vidové protějšky, na ortografické varianty svého lemmatu, případně také na protějšky v jiných slovnících).



Obrázek 4.7: Schématické znázornění transformace VALLEXu ve formátu B do formátu SAMR pro lexém *započítávat, započítat, započíst.*

Vlevo ve formátu B vidíme dva vidové protějšky, přičemž dokonavý (pf) se ještě dělí na dvě varianty. Následuje první LU, která je platná pro všechna tři lemmata bez omezení. Valenční rámec je sdílený, zatímco glosa a příklad se štěpí pro různé vidy. Druhá LU se vztahuje výhradně na jednu variantu dokonavého vidu, všechny informace se tedy týkají tohoto jediného lemmatu. Vpravo je lexém ve formátu SAMR „rozgenerovaný“ do více slovníkových hesel, které jsou provázány vzájemnými odkazy pro zachování původní informace. Z ukázky je patrné, že první způsob zápisu je výrazně úspornější, ovšem za cenu složitějšího strukturování informací.

Srovnání VALLEXu ve formátu B a ve formátu SAMR

Na závěr krátce porovnáme oba odlišné formáty pro VALLEX. Expandovaný formát SAMR je podle očekávání mnohem větší, z 8 MB a 200 000 řádek ve formátu B se stane 12 MB a 300 000 řádek ve formátu SAMR. Výhodou ovšem je mnohem pohodlnější práce při použití běžných unixových nástrojů. Uvedeme tři příklady hledání ve slovníku.

↪ *Příklad 1:* Kolik je ve slovníku iterativ?

V SAMRu je to jediný `grep` a spočítání vrácených řádek (`wc -l`):

```
grep '<word .*aspect="iter' vallex-2.5-samr.xml | wc -l
```

Kdežto v originálním formátu s něčím tak jednoduchým nevystačíme, potřebujeme nástroj pro zpracování XML. Ukážeme, jak dotaz vypadá například ve vyhledávacím a editovacím nástroji XSH:[76]

```
xsh -C 'open vallex-2.5.xml;
        regns v http://ufal.mff.cuni.cz/vallex-2.0;
        count //v:lexeme/v:lexical_forms//v:mlemma[@aspect="iter"]'
```

Dotaz je složitější proto, že element `mlemma` může být uveden buď přímo, nebo ještě zanořený do elementu `mlemma_variants`. Nesmíme ovšem počítat ty elementy, které patří lexikální jednotce a pouze upřesnění m-lemmata, která už jsme započítali jednou v záhlaví lexému.[77]

↪ *Příklad 2:* Kolik je ve slovníku lexikálních jednotek pro iterativa?

V SAMR formátu jsou to zkrátka LU patřící pod iterativní lemma. Tentokrát již také použijeme XSH:[78]

```
xsh -C 'open vallex-2.5-samr.xml;
        regns v http://ufal.mff.cuni.cz/vallex-samr;
        count //v:word[contains(@aspect, "iter")]//v:frame;'
```

Zjišťujeme počet elementů `frame` pod slovníkovými hesly obsahujícími `iter` v atributu `aspect`.

V původním formátu B je potřeba složitější výpočet:

[76] `http://xsh.sourceforge.net`

[77] Oba dotazy shodně vrací 307 iterativ ve VALLEXu 2.5.

[78] Nechme stranou možnost využít identifikátory, ve kterých „shodou okolností“ je vždy uveden vid daného lemmatu. Dotaz by potom byl takto jednoduchý

```
grep "<frame .*iter" vallex-2.5-samr.xml | wc -l
```

, nicméně identifikátor tuto vlastnost nezaručuje.

Počet LU pro iterativa -- dotaz v XSH

```
1 xsh -C '
2     open vallex-2.5.xml;
3     regns v http://ufal.mff.cuni.cz/vallex-2.0;
4     my $sum = 0;
5     foreach //v:lexeme/v:lexical_forms//v:mlemma[@aspect="iter"] {
6         if (not (@coindex)) cd ..;
7         $coindex = @coindex;
8         $count := count :q ../../v:lexical_units//v:blu[
9             count(v:lexical_forms)=0
10            or
11            v:lexical_forms/*[@coindex = $coindex]
12        ];
13        $sum = $sum + $count;
14    }
15    print $sum;'
```

Musíme již sčítat LU přes všechna iterativní m-lemmata (kterých může být víc v rámci jednoho lexému). Řádek 6 nás přesune na element `mlemma_variants`, pokud takový existuje. Potom si zjistíme `coindex` (7) a spočítáme všechny LU (`blu`), které nemají omezení na lexikální formy (tedy se týká všech lemmat v záhlaví, včetně zpracovávaného iterativa, 9), nebo mezi vyjmenovanými lexikálními formami je zpracovávané iterativum uvedeno (11). Na závěr sečtené počty vypíšeme (15).[79]

Jsou samozřejmě i dotazy, které je snazší zodpovědět ve formátu B. Dalo by se říci, že to jsou dotazy pracující s hlubší lingvistickou znalostí.

↪ *Příklad 3:* Zjisti počet m-lemmat (tedy zanedbej reflexivní částice) a započítej každý homografy zvlášť.

Ve formátu verze B to dokážeme snadno zjistit v prostředí unixového shellu. Najdeme všechny řádky obsahující `mlemma`, ponecháme z nich pouze část s m-lemmatem a případně s homografem a spočítáme unikátní řádky:

```
grep "<mlemma " vallex-2.5.xml \
  | grep -o "homograph.*\|'>.*" \
  | sort | uniq -c | wc -l
```

Ve formátu SAMR je to trochu složitější a navíc musíme od každého lemmatu odstranit případnou reflexivní částici:[80]

[79] Opět shodně oba dotazy vrátí stejný počet, 854 lexikálních jednotek.

[80] Výsledek vychází 3 553 m-lemmat.

```
grep "<lemma>\|homograph=" vallex-2.5-samr.xml \
  | sed -e '/homograph/N; s/^.*homograph="\(.\)".*\n/\1/' \
  | sed -e 's/ s[ei]//' \
  | sort | uniq -c | wc -l
```

Jiné možné řešení problému s formátem by mohlo být toto: nehleďme na formát, na něm nezáleží, důležité je efektivní a komplexní API k němu.

U tohoto přístupu vidíme dvě nevýhody, a proto jsme se touto cestou nevydali. Jak jsme viděli na předchozích stránkách, zachycovaná skutečnost je složitá v každém formátu a stejně složité mohou být dotazy na ni. Příprava API, které by slovník zpřístupňovalo, by tudíž byla nutně také dost složitá – pokud by se vůbec při jeho tvorbě na vše pamatovalo. Druhá výhrada spočívá v tom, že formát SAMR nám nejen zjednoduší přístup k VALLEXu, ale zajistí také přístup k PDT-Vallexu (ačkoli to není zadarmo a v některých aspektech se práce s PDT-Vallexem stále odlišuje). Bez něj bychom museli psát druhé API pro PDT-Vallex.

Od mnohem jednoduššího formátu SAMR si také slibujeme, že zpřístupní slovník VALLEX širší skupině jazykovědců mimo ÚFAL. Od verze 2.6 proto zveřejňujeme slovník v obou XML formátech.

4.14 Poznámka ke standardním formátům lingvistických dat

Nabízí se otázka, proč jsme pro účely porovnání (a prolinkování) nepoužili některé ze standardních a široce diskutovaných formátů pro zachycení lingvistických dat.

Takovými standardy jsou například ISO normy:

Linguistic Annotation Framework (LAF, ISO:24612:2012), (Eckart, 2012) je formát, který umožňuje reprezentovat jazykové anotace například korpusů, mluvené řeči či videa. Nabízí prostředky pro analýzu povrchového textu do tokenů (tedy základních elementů textu: slov a interpunkce), jejich morfologickou analýzu až po vystavění složkového syntaktického stromu. Data se serializují jako XML, nad kterým existuje abstraktní datový model.

Morpho-syntactic Annotation Framework (MAF, ISO:24611:2012) slouží k anotaci slovních tvarů textu. Jednotlivým tokenům přiřazuje lexikální jednotky (může odkazovat do slovníku) a morfo-syntaktické vlastnosti. Tyto vlastnosti se zapisují jako „struktura rysů“ („Feature Structure“) a je možné si definovat libovolné vlastnosti. MAF by tedy bylo možné použít spíše pro přiřazování lexikálních jednotek a valenčních rámců jednotlivým výskytům sloves v textu než k popisu celého slovníku odděleného od dat.

↪ *Příklad:* Uvedeme ukázku francouzského *belle* anotovaného v Morphosyntactit Annotation Framework.[81]

```
<token id="t0">belle</token>
<wordForm entry="urn:lexicon:fr:beau" lemma="beau" tokens="t0">
  <fs>
    <f name="pos"> <symbol value="adjective"/> </f>
    <f name="adj_type"> <symbol value="qualifier"/> </f>
    <f name="gender"> <symbol value="feminine"/> </f>
    <f name="number"> <symbol value="singular"/> </f>
  </fs>
</wordForm>
```

Zajímavostí použitého XML je, že umožňuje tzv. kompaktní reprezentaci. V ní se celý blok XML dokumentu nahradí jediným atributem, který nese stejnou informaci, spojenou do jediného textového řetězce. Dlouhá, nicméně srozumitelně strukturovaná informace o francouzském *belle* využívá vlastností, kvůli kterým bylo XML vytvořeno. V kompaktní reprezentaci ovšem MAF umožňuje přepsat ji i takto:

```
<token id="t0">belle</token>
<wordForm entry="urn:lexicon:fr:beau" lemma="beau" tokens="t0"
  tag="pos.adj adj_type.qual gender.fem num.sing"/>
```

Další standardy pro lingvistické zdroje jsou pak pro lexikografické účely nevhodné zcela.

Syntactic Annotation Framework (SynAF, ISO:24615:2010) (Declerck, 2006) je syntaktický příbuzný standardu MAF. Umožňuje syntaktickou anotaci textu, a to jak pomocí složkových, tak i závislostních stromových struktur. Pro účely lexikografické je tedy nevýhodný.

Semantic Annotation Framework (SemAF-Time, ISO:24617-1:2012, a DiAML, ISO:24617-2:2012) poskytuje způsob, jak standardizovaně zachytit časovou sémantiku textu a jak anotovat dialog.

Lexical Markup Framework (LMF, ISO:24613) (Francopoulo et al., 2006) nabízí metamodel pro reprezentaci jedno- i vícejazyčných lexikálních databází. Součástí LMF je také mechanismus integrace různorodých lexikálních zdrojů, tedy jejich vzájemného provazování. Je pravděpodobné, že by ho s použitím syntaktického a sémantického modulu a spolu s vlastními rozšířeními LMF bylo možné využít k zachycení značné části slovníků VALLEX a PDT-Vallex.

[81] Příklad je převzatý z `http://atoll.inria.fr/~clerger/MAF/html/body.1_div.10.html#body.1_div.10_div.2`.

4.14 POZNÁMKA KE STANDARDNÍM FORMÁTŮM LINGVISTICKÝCH DAT

Vraťme se tedy k otázce: Proč jsme LMF nebo jiný standardní univerzální formát nevyužili?

Je důležité zvážit, kolik by využití univerzálního formátu přineslo práce, a v čem by se to vyplatilo.

Zmíněné standardy jsou metamodely – v rámci nich by bylo potřeba teprve navrhnout model např. pro VALLEX. Základní, obvyklé prvky slovníků jsou v LMF k dispozici, jiné, specifické pro valenční slovníky, by bylo potřeba řešit novými rozšířeními, která bychom museli vytvořit.

Z pohledu našeho úkolu nám ovšem stačí mít formát univerzální jen do té míry, aby postihl VALLEX a PDT-Vallex. Ty byly vždy uloženy v relativně blízkých formátech a jejich úprava ke společnému formátu SAMR (viz sekce 4.13) je tedy zcela dostačující a poměrně snadná.

Jinak řečeno, univerzální formát pro tuto práci nepotřebujeme. Vytváření nového formátu pro naše slovníky, který by byl založený na některém univerzálním formátu, se míjí s prací, kterou zde potřebujeme skutečně provést. Pro samotnou práci se slovníkem je nakonec stejně více než formát důležité rozhraní, přes které se k němu (na programátorské bázi) přistupuje. Hledání společného formátu rozhodně není nejvýznamnější problém, který v souvislosti s prolinkováním řešíme (viz též Bejček, Kettnerová a Lopatková, 2014 a strana 118 v sekci Analýza chyb: pro provázání slovníků je podobná granularita jejich slovníkových hesel a podobnost informací, které lze k porovnání využít, mnohem důležitější než jejich datové formáty).

Zde je též vhodné připomenout, jaké formáty používají slovníky, které jsme zde představovali. Ani jediný z nich nepoužívá žádný z univerzálních formátů (alespoň nikoli jako svůj nativní – existují však projekty jako například UBY-LMF,[82] které různé slovníky konvertují do univerzálních formátů). Naprostá většina z nich[83] je uložena v XML, které je vytvořeno přímo pro potřeby daného slovníku. V takovém případě může být zvolen formát nejjednodušší a nejsrozumitelnější, zvláště srovnáme-li ho s formátem univerzálním, přizpůsobeným až následně pro jeden konkrétní slovník.

[82] `https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/`

[83] Jmenovitě VerbaLex, Český WordNet, FrameNet, slovník PropBanku i NomBanku, VerbNet a také dvojstranné prolinkování slovníků v rámci SemLinku.

Část II

Propojování jazykových zdrojů

5

Propojování slovníků mezi sebou

Klíčovou část této práce tvoří propojování slovníků mezi sebou. V dnešní době existuje značné množství nejrůznějších elektronických – tedy strojově snadno zpracovatelných – slovníků (některé jsme představili v kapitole 4), a to i pro češtinu, dokonce i pro oblast valenční. Jejich oddělené, samostatné používání je samozřejmě možné a v historii tomu tak také vždy bylo. Dnes však již z technického pohledu není důvod nemít více slovníků propojených. Takové propojení pak například z uživatelského hlediska umožní jedním dotazem získat informace ze všech relevantních slovníků najednou (ať už v nějaké agregované formě, nebo odděleně po jednotlivých slovnících).

↪ *Příklad:* Takový přístup ukazuje například Internetová jazyková příručka Ústavu pro jazyk český Akademie věd ČR.[1] Pokud jsme v ní vyhledali např. slovo *kytovec*, přinášela vždy pravopisné informace („rod: m. živ., …“) a tabulku s rozpisem celého paradigmatu („1. pád kytovec, 2. pád kytovce, 3. pád kytovci/kytovcovi, …“). V poslední době však přibyly také odkazy do výkladových SSČ, SSJČ a Akademického slovníku cizích slov (kde zjistíme, že jde o „mořského savce jen s předními končetinami ve tvaru ploutve“, či že slovo pochází z ruštiny).

Zmíněné slovníky jsou propojeny jen na úrovni slovníkových hesel, tedy podle lemmat, a jen výjimečně přítomná ambiguita se nijak neřeší. Propojovat se ale dají i jiné typy slovníků, na hlubší úrovni (třeba jednotlivé významy výkladových slovníků), nebo napříč jazyky.

Každým takovým spojením dochází jednak ve výsledku k rozšíření rozsahu informací podávaných každým ze slovníků, jednak k porovnání různých lingvistických teorií, na jejichž základě slovníky vznikaly.

Pro angličtinu je propojování lexikálních databází a slovníků dlouhodobým trendem, jak jsme popisovali na příkladu propojení PropBanku, VerbNetu, FrameNetu a WordNetu v sekci 4.8 SemLink.

V následujících sekcích popíšeme propojování těchto dvojic slovníků:

[1] `http://prirucka.ujc.cas.cz`

5.1 VALLEX + PDT-Vallex – Propojení dvou mírně odlišných českých valenčních slovníků. Motivací je jak rozšíření informací pro PDT-Vallex, tak i napojení VALLEXu na data PDT.

5.2 VALLEX 1.0 + VALLEX 2.5 – Propojení dvou verzí stejného českého valenčního slovníku. Motivací je získání přístupu k datům VALEVALu pro novější VALLEX.

5.3 VALLEX + FrameNet – Propojení dvou odlišných valenčních teorií, navíc mezi češtinou a angličtinou. Motivací je doplnění sémantické informace do VALLEXu.

5.4 PDT-Vallex + FrameNet – Tranzitivní propojení přes VALLEX. Motivací je anotace PDT i PCEDT sémantickými rámci z FrameNetu.

5.1 VALLEX a PDT-Vallex

S propojováním začneme hned nejdůležitější dvojicí slovníků, které jsme představili v sekcích 4.2 a 4.3. Propojování probíhalo v hlavním projektu této práce, který nazýváme Vallink. Vysvětlíme, proč existují dva podobné valenční slovníky, které je nyní potřeba propojit (5.1.1), a jaké výhody nám to přinese (5.1.2). Upřesníme úlohu, kterou se budeme zabývat (5.1.3) a přejdeme k jejímu řešení (5.1.4) pomocí porovnávání širokého spektra informací z jednotlivých lexikálních jednotek/valenčních rámců slovníků. Následně postup vyhodnotíme (5.1.5), představíme praktické výstupy z projektu, které jsou zveřejněny (5.1.6) a upozorníme na problémy, kterých jsme si vědomi (5.1.7). Dotkneme se opakovaně aplikovaného provazování slovníků (5.1.8) a sekci uzavřeme (5.1.9).

Projekt popsaný v této sekci byl zveřejněn ve stručné formě jako Bejček et al. (2014).

Dříve, než jsme provázali slovníky VALLEX a PDT-Vallex, vyzkoušeli jsme si postup na jednodušším případě dvou verzí stejného slovníku, o čemž zde ovšem pojednáme až následně v sekci 5.2.

5.1.1 Historie

Nebude na škodu, pokud problematiku otevřeme stručným vysvětlením, jak vůbec došlo k situaci, kdy na jednom pracovišti vznikly dva *valenční* slovníky pro *češtinu*. Reálné plány na sestavení valenčního slovníku českých sloves sahají k počátku tohoto století – teoretické plány tu však byly od počátku Funkčního generativního popisu (FGP). Od jara 2001 začal v Ústavu formální a aplikované lingvistiky vznikat elektronický slovník VALLEX. Ovšem roku 2000 byl dokončen Pražský závislostní korpus (PDT) verze 1.0 (Hajič et al., 2001) a součástí nové verze 2.0

měla být také anotace valence (sloves a některých substantiv, adjektiv a adverbií). Aby bylo dosaženo kvalitní a konzistentní anotace (tedy aby např. stejná lexikální jednotka měla zaručeně vždy tatáž obligatorní doplnění), bylo potřeba mít valenční slovník. VALLEX však v tu dobu ještě neexistoval, přinejmenším ne v dostatečném rozsahu. Proto začal paralelně s VALLEXem vznikat druhý valenční slovník PDT-Vallex. Zatímco VALLEX zabíhal „do hloubky“ a doplňoval další informace k vybraným slovesům, PDT-Vallex se musel rozběhnout „do šíře“ a pokrýt všechna slovesa vyskytnuvší se v PDT. Vznikající VALLEX tak nebylo možné použít v průběhu editace ke kontrole valence v PDT.

Oba slovníky představené v sekcích 4.2 a 4.3 se liší výběrem slovesných lemmat, rozsahem i konkrétní aplikací teoretického přístupu. Shrňme zde, v čem odlišnosti spočívají.

Výběr sloves

Nový záznam byl do PDT-Vallexu vložen ve chvíli, kdy se lexikální jednotka při anotaci PDT poprvé objevila v datech. Pokud to byla první lexikální jednotka slovesa, teprve potom bylo založeno nové slovesné heslo (v opačném případě jen přibyl nový valenční rámec pod již existující heslo). Slovesa pro VALLEX byla vybírána podle četnosti v (části) SYN 2000 a poté byla zpracována kompletně, s ambicí pokrýt všechny lexikální jednotky.[2]

Rozsah slovníku

VALLEX 2.6 čítá 4 790 sloves dělících se v průměru na 2,3 lexikálních jednotek. PDT-Vallex 2.0 obsahuje 7 100 sloves, ale průměrně jsou popsána pouze 1,7 lexikálními jednotkami. (Viz tabulka 5.1.) Na druhou stranu, zatímco VALLEX si neklade za cíl úplný popis všech idiomů, v PDT-Vallexu se na mnohé z těchto nepopsaných obratů při anotaci narazilo, proto není výjimečné, že konkrétní sloveso má mnohem víc lexikálních jednotek v PDT-Vallexu než ve VALLEXu.[3] Kromě sloves pak PDT-Vallex obsahuje také 4 530 substantiv, adjektiv a adverbií.

[2] Přesněji všechny neidiomatické lexikální jednotky. Idiomy jsou zachycovány také, včetně příznaku, že se jedná o idiom, ale bez nároku na úplnost.

[3] Například *jít* s poměrem 59:15 lexikálních jednotek ve prospěch PDT-Vallexu. Z následujících dvaceti frází v PDT-Vallexu jsou ve VALLEXu jen tři (označené hvězdičkou): *„jít s něčím dohromady“**, *„jít k duhu“*, *„jít někomu po krku“*, *„jít pod vousy“*, *„jít s někým s kopce“*, *„jít s cenou [nahoru/dolů]“*, *„jít se klouzat“*, *„jít na odbyt“*, *„jít příkladem“**, *„jít proti proudu“*, *„jít s tím ruku v ruce“*, *„jít vzorem“*, *„jít s kůží na trh“*, *„jít do tuhého“**, *„jít tlustý do tenkejch“*, *„jít do háje“*, *„jít ke dnu“*, *„jít na nervy“*, *„jít od válu“*, *„jít proti srsti“*.

Dále uvádí (už jako nefrazeologické) třeba valenční rámce se směrovým doplněním, pro něž

	VALLEX ver. 2.6	PDT-Vallex ver. 2.0
Počet různých slovesných lemmat	4 787	7 103
Průměrný počet LU na lemma	2,30	1,68
Celkový počet lemmat ve slovníku	4 887	11 656

Tabulka 5.1: Počet lemmat a lexikálních jednotek ve VALLEXu a PDT-Vallexu. Toto je jen výběr z tabulky 5.2 na straně 109, která obsahuje více statistických údajů a u níž je také podrobný popis.

Praktická aplikace teorie

Po teoretické stránce jsou oba slovníky v zásadě totožné (neboť oba mají stejný teoretický základ FGP a používají stejné pojmy), ale pravidla a postupy výroby hesel zcela stejné nebyly. Systematicky se lišily přístupy autorských kolektivů k zachycení některých konkrétních jevů, například:

analytické predikáty: Ve VALLEXu nejsou dosud (ve verzi 2.6) zachycované analytické predikáty. Zato v PDT-Vallexu namnoze dochází k oddělení analytických predikátů od syntakticky totožné lexikální jednotky. Pro analytické predikáty je vytvořena druhá lexikální jednotka určená otevřeným výčtem, jako je tomu v případě slovesa *pustit se* v PDT-Vallexu:
ACT(1) PAT(do+2) a
ACT(1) CPHR(do-1[{akce,boj,bojkot,čtení,hospodaření,hra,investice, obchodování,podnikání,polemika,práce,projekt,příprava,psaní,řešení, spekulace,testování,výklad,výroba,vývoj,výzva,…}.2]),
zatímco ve VALLEXu je vše sloučeno pod první tvar rámce, dokonce s výslovným příkladem *„pustit se do práce“*. Navíc tam pak přibývá stejně vypadající LU pro idiom *„pustit se do někoho“* s významem *zaútočit*, či *dobírat si*, který v PDT-Vallexu celkem pochopitelně chybí (není to idiom obvyklý pro novinové texty). Pokud tento případ shrneme, oba slovníky mají dvě LU se shodnou morfematickou realizací, které si ale vzájemně neodpovída-

také ve VALLEXu nenacházíme ekvivalenty, případně spadají do doslovného rámce *„přemisťovat se chůzí“* ACT(1) MANN^{typ} ↑DIR^{typ}. Obraty, které těmto valenčním rámcům odpovídají, jsou například: *„jde to z jeho srdce“*, *„jít v jeho šlépějích“*, *„jít do dražby“*, *„jít sám na brankáře“*, *„jdou na to desetiny procenta“*, *„náklady jdou do statisíců“*, *„kam až jde ve svých úvahách“*.

jí, což je způsobené odlišným pojetím delimitace významů do lexikálních jednotek.

delimitace do LU: Zmíněná odlišná delimitace je širší problém, nesouvisí jenom s analytickými predikáty. Oba slovníky se však ve vymezení lexikálních jednotek liší velmi často, což se odráží v odlišném počtu LU pro odpovídající si lexémy. Ať už je důvodem neúplné pokrytí idiomů, frází a analytických predikátů ve VALLEXu (sloveso *mít* má 135 LU v PDT-Vallexu a jen 23 ve VALLEXu), nebo jiná jemnost dělení významů (PDT-Vallex vyděluje zvláštní lexikální jednotky pro schopnost, např. *„Jeník už čte“* je $číst^P_3$, ale ve VALLEXu to patří pod obecné $číst^V_1$).[4]

výběr funktorů: Samotná sada používaných funktorů se liší. Většina je sice společná oběma, ve VALLEXu však byly navíc používány funktory OBST a RCMP, kdežto PDT-Vallex používal navíc CPHR, CPR, RESL, RESTR, THO, TOWHN a TPAR (ačkoli s výjimkou prvních dvou se nikdy nevyskytly jako obligatorní, pročež jsou uváděny pouze v příkladech).

zachycení fakultativních volných doplnění: Funktor je v PDT-Vallexu uveden přímo v rámci, pouze pokud je to aktant a/nebo je obligatorní. Ostatní funktory typické pro danou LU se vyskytují nanejvýš u příkladů. Vazba *„léčit někomu ránu mastí“* tak má ve VALLEXu rámec ACT(1) PAT(4) $\mathsf{MEANS}(7)^{\text{typ}}$ $\mathsf{BEN}(3)^{\text{typ}}$, ale v PDT-Vallexu jsou fakultativní volná doplnění vynechány: ACT(1) PAT(4), jen funktor MEANS je pak uveden v příkladu: *„léčit bezlepkovou dietou.MEANS“*.

zápis morfematických vyjádření: Oba slovníky se v zachycení morfematických vyjádření jednotlivých doplnění podstatně liší. PDT-Vallex zachycuje doplnění na analytické a morfologické rovině, tedy jako závislostní vztahy mezi slovy, která jsou reprezentována lemmatem plus hodnotami relevantních morfologických kategorií, zatímco VALLEX se spokojí s pouhým zápisem předložek a pádů,[5] nebo v případě frazémů s konkrétními slovními formami. Na frazému, jako je *„postavit se na vlastní nohy“*, to snadno demonstrujeme: v obou slovnících jde shodně o ACT DPHR, ale zatímco VALLEX uvádí pouze DPHR(na vlastní nohy), PDT-Vallex upřesňuje DPHR(na-1[noha:P4[vlastní-1.#]]), tedy že *noha* je v akuzativu plurálu a závisí na předložce *na* (kde „-1“ je desambiguační sufix, který odlišu-

[4] Pro stručnější a přehlednější popis zaveďme značení, kde lemma s horním indexem reprezentuje slovníkové heslo, tedy celý lexém (V pro VALLEX a P pro PDT-Vallex) a s dolním indexem reprezentuje konkrétní lexikální jednotku. Tedy $plést\ se^V_4$ je čtvrtá LU slovesa *plést se* ve VALLEXu (s významem *„mýlit se“*) a tato zrovna odpovídá LU $plést\ se^P_3$.

[5] Obvykle se jedná o pád, či pád s předložkou. Dále se však zachycují také vedlejší věty se spojkami i bez nich, infinitivní vazby, adjektiva, apod., viz strana 27.

je předložku od citoslovce) a že *vlastní* závisí na *noha* a je s ní ve shodě v pádě a čísle (znak ‚#').

zpracování frází: Spojení *„postavit někoho mimo službu"* je ve VALLEXu chápáno jako idiom, tedy ACT(1) PAT(4) DPHR(mimo službu), kdežto v PDT-Vallexu je to zachyceno čistě syntakticky jako směr někam (ve významu stavovém), spolu s např. *„postavit do reálného světla"*, tedy ACT(1) PAT(4) DIR3(=), kde ‚=' označuje stavový význam.

množství informací: Zatímco PDT-Vallex má, jak jsme už zmínili, detailnější popis morfematických vyjádření, VALLEX naopak lexikálním jednotkám systematicky připisuje další syntaktické (a syntakticko-sémantické) informace (deagentní diateze, reflexivita, reciprocita, kontrola, syntakticko-sémantická třída, idiomaticita), které v PDT-Vallexu nejsou.

jednorázové odlišnosti: Krom systematických rozdílů je zde také celá řada rozdílů nepravidelných. Ty jsou způsobené jiným pohledem autorů na danou LU. Stejně jako mezi dvěma anotátory budou neshody, liší se i pohled obou autorských týmů například na obligatornost některých konkrétních doplnění, nebo zda je vhodné jemné rozdíly ve významu sloučit do jediné lexikální jednotky.[6] Několik příkladů jednotlivých nesystematických rozdílů:

- Idiom *„nalézt klíč k (řešení)"* je zachycen v PDT-Vallexu, ale opominut ve VALLEXu.
- Lexikální jednotka *„myslet tím něco"* má prohozené funktory:
  ACT(1) PAT(4,že) EFF(7) ve VALLEXu a
  ACT(1) EFF(4,že) PAT(7,pod+7) v PDT-Vallexu. (Pro jednoduchost zde sjednocujeme zápis rámců obou slovníků zhruba podle VALLEXu.)
- Někdy je stejné valenční doplnění v obou slovnících označeno jiným funktorem. Například functor BEN je ve VALLEXu často používán i pro doplnění, která jsou v PDT-Vallexu značena jinak, jako ADDR, nebo PAT. Ale takových „záměn" je více, např.:
  - *„mávat někomu"*: BEN vs. ADDR
  - *„mávat na někoho"*: AIM vs. ADDR
  - *„hnout žlučí někomu"*: BEN vs. PAT
  - *„krýt před něčím"*: EFF vs. ADDR

[6] V obou slovnících je sice hlavní kritérium pro delimitaci syntaktické a v obou také platí, že rozhodování o vyčlenění nové LU bylo řízeno primárně na základě formálních znaků, ani v jednom však nebylo formální hledisko to jediné. Ve VALLEXu docházelo častěji k rozdělení na dvě LU se stejným valenčním rámcem, pokud se významně lišil význam obou jednotek.

5.1.2 Motivace

Když jsme připomněli, čím se slovníky liší, je dobré se podívat, co můžeme získat jejich propojením. Obecně: oba získají právě to, čím se od nich druhý slovník liší a co jim dosud chybělo.

Vzroste **množství zpracovaných sloves i slovesných lexikálních jednotek** ve srovnání s jedním i druhým slovníkem. Ani jeden slovník totiž není podmnožinou toho druhého: VALLEX obsahuje častá slovesa, která se ale nemusela nutně objevit v novinových textech PDT, a také pro úplnost obsahuje vidové protějšky častých sloves i v případě, že se tyto protějšky vyskytují zřídka; PDT-Vallex zase popisuje celou řadu sloves, která se vyskytují řidčeji a do frekvenčního výběru pro VALLEX se již nemohla dostat, přesto se však v novinách používají a vyskytla se také (alespoň jednou) v PDT. PDT-Vallex má pak z výše uvedených důvodů nižší počet lexikálních jednotek na jedno sloveso, proto spojením s VALLEXem získá (jen pro společná slovesa) úplnější popis všech možných LU.[7] Ale i v PDT-Vallexu se nalezne LU, která nemá ve VALLEXu obdobu (stává se zejména u frazeologie) a měla by být doplněna.

VALLEX je dosud jen nedostatečně propojen se skutečnými užitími sloves ve větách v korpusu. Pro zhruba 200 sloves jsme získali průměrně téměř padesát anotovaných vět, jak bude popsáno v sekci 6.4 PDT/PCEDT a VALLEX. To je pro slovník velikosti VALLEXu nepřiměřeně málo. Propojením s PDT-Vallexem ovšem získáme **propojení LU s větami v PDT a v PCEDT**, a tedy řadu korpusových příkladů.

PDT-Vallex na druhé straně zachycuje jen nejnutnější informace, které byly potřebné při anotaci PDT:

- počet a druh obligatorních a fakultativních valenčních doplnění,
- povrchovou realizaci všech obligatorních i fakultativních doplnění včetně syntaktické struktury,
- glosu a příklad a
- pomocné informace (dosavadní počet výskytů dané LU v datech).

K nim PDT-Vallex po propojení získá další syntaktické informace z VALLEXu, jako jsou údaje o možnosti a typu reflexivizace, reciprocity a kontroly. Dále též syntakticko-sémantické informace, jako jsou vid, explicitní označení idiomu a příslušnost k syntakticko-sémantické slovesné třídě. PDT-Vallex také nijak neřešil

[7] Úplný v tom smyslu, že v korpusových a slovníkových zdrojích se žádná další LU nevyskytla. Neznamená to, že v budoucnu nemůže být doplněna nějaká opomenutá LU, která se nevyskytuje ani ve starších slovnících, ze kterých se vycházelo při tvorbě VALLEXu, viz str. 26. Nejčastěji to bude případ idiomatického užití či výrazu z jiné domény, což naznačuje i výrazné rozšiřování PDT-Vallexu při pokračujících anotacích textů například v české části PCEDT.

homografy[8] na jedné straně (ty jsou ve VALLEXu odděleny) ani vidové protějšky a ortografické varianty jednoho slovesa na straně druhé (ty jsou ve VALLEXu všechny vyjmenovány a propojeny). Přes VALLEX se tedy oddělí například *dovést-I* se dvěma LU od *dovést-II* s jedinou LU. Naopak se propojí například slovesa *myslet* a *myslit* a také $nahlížet^{impf}$ a $nahlédnout^{pf}$.

Stručně shrnuto, oba slovníky rozšíří počet hesel a LU, PDT-Vallex získá doplňující informace k některým svým heslům a VALLEX získá odkazy od některých svých hesel do dat.

5.1.3 Popis úlohy

Naší úlohou tedy je provázat hesla z PDT-Vallexu s hesly pro stejná slovesa ve VALLEXu a naopak. V rámci takto identifikovaných stejných sloves pak provázat jednotlivé lexikální jednotky z jednoho slovníku s odpovídajícími[9] lexikálními jednotkami ve slovníku druhém.

Cílem však není vznik jediného slovníku, naopak, oba se mohou nadále vyvíjet odděleně, každý navázaný na další projekty. Důvodem je zejména odlišný přístup při tvorbě obou slovníků a jiná pravidla pro zachycení některých jevů. Proto je nutné provazování hesel navrhnout tak, aby je bylo možné v budoucnu provádět opakovaně (polo) automaticky. Pro uživatele slovníků by to měla být doplňující informace a fakt, že pochází z jiného slovníku, jim může být zcela skryt. Uživatelé by měli mít možnost s oběma slovníky pracovat najednou, vyhledávat ve sjednoceném seznamu sloves, moci jednoduše přecházet při vyhledávání chybějící informace z jednoho slovníku do druhého, zobrazovat snadno věty z PDT pro lexikální jednotky ve VALLEXu atp.

---

[8] Jako *homografy* označujeme ty významové varianty slovesa, které mají stejnou grafickou podobnu, nicméně není zřejmé, že by spolu souvisely sémanticky. Jsou pro ně tedy vyhrazeny různé lexémy, které mají stejné lemma a jsou doprovozeny římskou číslicí. Homografy byly v PDT-Vallexu dosud pro jednoduchost spojeny v jediném hesle.
↪ *Příklad: dovést-I* ve smyslu *„přivést, doprovodit“* a *dovést-II* ve smyslu *„umět, dokázat“* se významově i syntakticky výrazně liší: *dovést-II* se považuje za modální sloveso, (viz Mikulová et al., 2006, oddíl 6.9.1.1), netvoří trpný rod, substantivní derivativa od obou slov jsou odlišná: *dovedení* vs. *dovednost* a v SSJČ jsou uváděny jako dvě hesla. Ovšem v PDT-Vallexu jsou všechny významy spojeny v jediném slovníkovém hesle $dovést^{P}$ se třemi rámci.

[9] Budeme říkat, že dvě LU z různých slovníku si vzájemně *odpovídají*, když vyjadřují natolik stejný význam, že pokud by se obě objevily v jediném slovníku, nebyly by rozděleny do dvou jednotek. Takové rozhodnutí je pochopitelně subjektivní a je věcí autora slovníku. Ale stejně tak je věcí konsenzu lexikografů, případně věcí rozhodnutí vybrané skupiny či jednotlivce ze své podstaty celá úloha propojování slovníků. Proto považujeme za korektní také o *odpovídajících si jednotkách* hovořit v tomto směru vágně, neboť nemůže existovat žádná přesná definice.

Na konci sekce 5.1.1 jsme uvedli některé drobné, leč významné odlišnosti VALLEXu a PDT-Vallexu. Šlo například o odlišné zachycení frází, jiné funktory pro popis stejných rámců, specifikaci struktury valenčních doplnění, ale i o jednorázové odlišnosti dané jiným jazykovým citem autorů a zejména o odlišnou delimitaci jednotlivých lexikálních jednotek.

Z těchto odlišností je již patrné, že provázání jednotlivých lexikálních jednotek mezi slovníky není úloha triviální a nemůže probíhat přímočaře podle jednoduchých pravidel (pokud by vůbec mohla probíhat plně automaticky). Vezměme si jedno obecné sloveso, které je popsané oběma slovníky. Předpokládejme, že na úrovni slovníkových hesel už jsou slovníky propojené a chceme provázat lexikální jednotky tohoto slovesa. Z obou slovníků je potřeba nejprve získat veškerou použitelnou informaci pro jednotlivé LU, a to navíc v porovnatelné formě. Na základě ní se pak (polo)automaticky rozhodnout, které LU si vzájemně odpovídají, a propojit je. Propojením se zde myslí jednoznačný odkaz od LU v jednom slovníku k LU druhého slovníku.

Není navíc zaručeno, že ke každé LU najdeme odpovídající protějšek. Různý počet LU (daný odlišnou delimitací významů) přímo implikuje případy, kdy jedné LU bude přiřazeno více LU v druhém slovníku, nebo jí naopak odpovídající LU zcela chybí.

Formálně vzato provázání $\mathfrak{m}$ lexikálních jednotek z VALLEXu s $\mathfrak{n}$ lexikálními jednotkami v PDT-Vallexu odpovídá přiřazení $0-\mathfrak{n}$ jednotek každé LU ve VALLEXu a $0-\mathfrak{m}$ jednotek každé LU v PDT-Vallexu, neboli nalezení až $\mathfrak{mn}$ dvojic lexikálních jednotek.

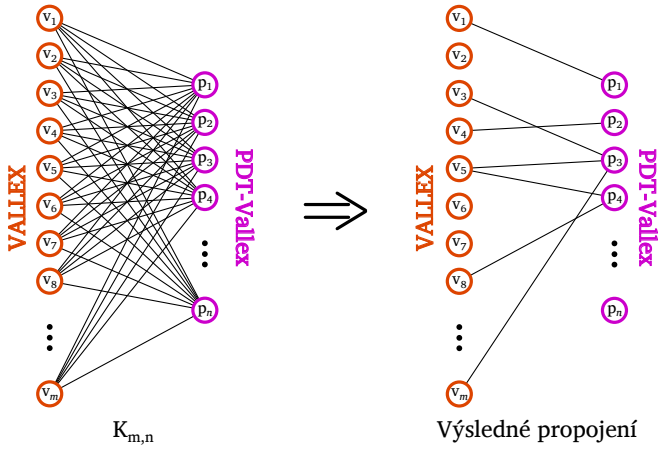
Užitečný je pohled na naši úlohu z hlediska teorie grafů. Máme dvě množiny vrcholů: V_V reprezentující LU VALLEXu, kde $|V_V| = \mathfrak{m}$ a V_P reprezentující LU PDT-Vallexu, kde $|V_P| = \mathfrak{n}$. Hledáme takový bipartitní graf, kde hrana spojuje vrcholy vzájemně si odpovídajících lexikálních jednotek. A priori není vyloučená žádná hrana, takže hledáme vhodný podgraf úplného bipartitního grafu $K_{\mathfrak{m},\mathfrak{n}}$, viz obrázek 5.1 vlevo. Podle lingvistických kritérií vhodný podgraf ovšem nemusí být ani *hranové pokrytí*[10] (neboť je-li v jednom slovníku zcela opominuta nějaká LU, pak odpovídající LU ve druhém slovníku musí zůstat nespárovaná, jako izolovaný vrchol, jak je vidět na obrázku vpravo pro LU v_6, či LU $p_{\mathfrak{n}}$), ani *párování*[11] (neboť jedna LU může odpovídat více LU v druhém slovníku, tak jako LU p_3 na obrázku odpovídá lexikálním jednotkám v_3, v_5 a $v_{\mathfrak{m}}$).

Ačkoli ani jedna vlastnost zaručena není, jsou obě žádoucí a jejich spojení, tzv. *perfektní párování*,[12] tedy vzájemné mapování 1:1, by pochopitelně bylo ide-

[10] *Hranové pokrytí* je taková podmnožina hran, že do každého vrcholu vede hrana.

[11] *Párování* je taková podmnožina hran, že žádné dvě nesdílejí stejný vrchol.

[12] *Perfektní párování* je párování pokrývající všechny vrcholy grafu.



Obrázek 5.1: Ukázka možného propojení lexikálních jednotek mezi VALLEXem a PDT-Vallexem. Vlevo úplný bipartitní graf, z něhož musíme vybrat ty správné hrany. Vpravo je jeden možný výsledek. Ve VALLEXu, kde je víc LU, mohou některé zůstat nespárované; ale i v PDT-Vallexu se může stát, že vhodný ekvivalent v druhém slovníku chybí. Naopak v případě odlišného dělení významů do LU dochází ke spojování jedné LU s více protějšky.

ální pokaždé, když to slovníky pro dané sloveso umožňují. Nejen že je v takové situaci snazší provést mapování automaticky, ale hlavně jsou hrany jako (v_4, p_2) ideální z pohledu uživatele budoucích propojených slovníků: od jedné LU se dostane *jednoznačně* k ekvivalentu v druhém slovníku a nemusí se zabývat žádným větvením možností.

5.1.4 Zvolené řešení

Náš postup provazování zhuštěný do šesti vět vypadá následovně: Nejprve najdeme ta slovesa, která jsou zpracována v obou slovnících. Tyto dvojice sloves porovnáváme ve více krocích. Nejprve porovnáme každou dvojici lexikálních jednotek, zda si může odpovídat a měla by být propojena. Tyto dvojice ohodnotíme číselným *Score*, které vznikne součtem ohodnocení podle dílčích kritérií. Dvojice, jejichž *Score* nepřesáhne určený práh, vyřadíme. Ze zbylých dvojic vybereme takové, které nejlépe pokrývají všechny lexikální jednotky. V této sekci popíšeme uvedený postup podrobně.

Společná slovesa

Nutně musíme začít nalezením sloves, jejichž lexikální jednotky je třeba provázat. Oba slovníky uvádějí u sloves lemmata, takže tato část je snadná. Jediný problém tvoří homografy (jako například *lekat*, viz poznámka 8 v sekci 5.1.2), které jsou ve VALLEXu rozděleny na dvě samostatná hesla rozlišená číselným sufixem. Avšak vzhledem k tomu, že v PDT-Vallexu homografy nijak odlišené nejsou, nemáme na úrovni lexému jinou možnost, než alespoň prozatím spojit oba homografy s jediným heslem v PDT-Vallexu a rozhodnutí nechat na úroveň lexikálních jednotek. (Tedy spojíme *lekat-I*^{V} s *lekat*^{P} a stejně tak spojíme i *lekat-II*^{V} s tímtéž *lekat*^{P}. To jak v případě, že heslo v PDT-Vallexu obsahuje jen význam *„někoho postrašit“*, nebo jen *„umírat na suchu“*, tak v případě obou významů přítomných v hesle současně jako dva různé rámce.) Spoléháme na to, že v následném provazování LU homografického slovesa se jeden z lexémů (žádná z jeho LU) se slovesem v PDT-Vallexu nespojí a v takovém případě dodatečně zrušíme i propojení na úrovni slovníkových hesel.

Problém samozřejmě je, že lexém ve VALLEXu je tvořen více lemmaty a ne každému lemmatu pak přísluší všechny lexikální jednotky (viz sekce 4.2 VALLEX). Tento a další problémy řeší společný formát, do kterého jsme slovník převedli v sekci 4.13 Formát SAMR pro valenční slovníky. Zde tedy chceme jen připomenout, že ani spojení na úrovni lexémů či lemmat není tak přímočaré, jak by se mohlo zdát, neboť v PDT-Vallexu nejsou vidové protějšky nijak shlukovány, aby se na ně dalo odkázat jako na celek, a ve VALLEXu zase není dobrý způsob, jak odkázat jen na část lexému odpovídající jednomu z lemmat.[13] Dál se tím zde nebudeme zabývat a předpokládáme, že je to zcela vyřešeno transformací do společného formátu SAMR.

Máme tedy připraveny dvojice slovníkových hesel, každá dvojice popisuje totéž sloveso. V této dvojici máme na jedné straně $\mathfrak{m}$ lexikálních jednotek, na něž se sloveso rozpadá ve VALLEXu, a na druhé straně $\mathfrak{n}$ rámců v PDT-Vallexu. Pokud si některé z nich vzájemně odpovídají, tedy pokud vyjadřují v zásadě tentýž význam, cílem je provázat je automaticky vzájemnými odkazy.

Obecně vzato, automatická procedura propojující dva slovníky se nemůže opřít o nic jiného než o informace obsažené v obou slovnících, a tudíž vzájemně porovnatelné. Čím větší je průnik obou slovníků v typech zachycovaných informací, tím snazší je úloha a tím větší bude úspěšnost kvalitního propojení.

Je tedy potřeba provést analýzu, které informace jsou ve slovnících společné a jakým způsobem jsou v nich zachyceny. Nejde jen o to, jak je technicky vy-

[13] Abychom byli úplně přesní, lze použít identifikátory přiřazené lexému a zkombinovat ho s označením lemmatu, které je v rámci lexému jednoznačné. Přijdeme však o možnost využít reference poskytované přímo v XML (DTD typy `ID` a `IDREF`).

řešeno zachycení, tedy jaký formát dat je zvolen pro uložení strojově čitelných informací ve slovníku. (Uložení informace bude obecně, s odhlédnutím od našich slovníků, jistě odlišné, buď i na vyšší úrovni – představme si například použití MySQL databáze versus souboru ve formátu YAML – nebo přinejmenším na nižší úrovni, jako v našem případě, kdy sice jde o dva XML soubory, ale lišící se vnitřní strukturou, pojmenováním a zanořením elementů, apod.) Kromě formátu je však také důležité porovnat, jak jsou dva stejné jazykové jevy zachyceny už na teoretické rovině: zda se stejný jev nenazývá pokaždé jinak, nebo není formalizován odlišně, zda zdánlivě neobsaženou informaci nedokážeme vyvodit z jiné, zda dokážeme využít neúplný popis nějakého jevu, atp.

V našem případě jsme našli tyto čtyři informace, které jsou společné oběma slovníkům a které by bylo možné využít pro porovnání LU. V následujícím textu je rozebereme podrobněji:

- valenční rámec,
- množina lemmat z příkladů a glos,
- reciprocita a
- kontrola.

Tyto i další informace jsou v nativních XML formátech pro oba slovníky kódovány značně odlišně, jak jsme ukázali v sekci 4.12. Společný formát SAMR jejich zápis sjednocuje alespoň do té míry, do jaké je to možné.[14] Tento formát jsme probrali v sekci 4.13 a dál nebudeme na odlišnosti upozorňovat. Předpokládáme, že ve formátu SAMR jsou slovníky dostatečně připraveny pro všechny operace, které bude potřeba nad slovníky provádět. Výjimku tvoří zápis typických funktorů, o nichž bude vhodnější napsat níže (str. 100).

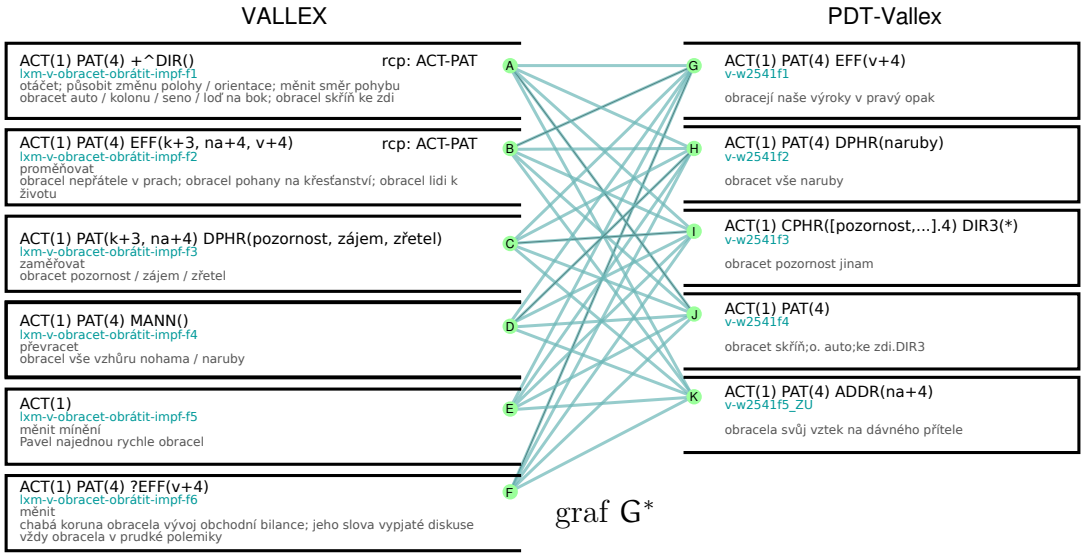
Porovnávání dvojic LU

Nyní tedy každou možnou dvojici lexikálních jednotek, označme ji (v_i, p_j), posoudíme podle podobnosti a přiřadíme jí *Score*: postupně ji ohodnotíme ve více kritériích na základě údajů, které jsou v obou slovnících společné, a za každé kritérium přiřadíme číselnou hodnotu; *Score* je potom tvořeno součtem těchto

[14] Například zápis složitějších morfematických forem je nepřevoditelný, proto je ponechán v původním tvaru. Ve většině případů by sice bylo možné zjednodušit závislostní zápis z PDT-Vallexu s využitím morfologických značek na doslovnou frázi použitou ve VALLEXu (například pro sloveso *chovat* nahradit „jako[v-1[bavlnka.S6]]“ na „jako v bavlnce“; příklad z PDT-Vallexu uvádíme v čitelnějším zápisu, který je s jeho XML formátem vzájemně jednoznačně převoditelný). Tím bychom ovšem ztratili podstatnou část informace a v případě složitějších závislostních vztahů není linearizace (zejména pořadí slov) jednoduchá úloha, viz *sedět* „na-1[židle.P6[dva.#]]“. Je zřejmé, že v důsledku toho se při porovnávání valenčních rámců takto zapsaná forma nemůže podobat formě ve VALLEXu.

ohodnocení. Čím vyšší hodnota, tím podobnější si dvě LU jsou. Jinou přesnější interpretaci *Score* nemá.

Postup znázorníme na slovese obracet, jehož LU jsou uvedeny na obrázku 5.2. Uprostřed obrázku je úplný bipartitní graf $K_{6,5}$, který si označíme $G^* = (V, E^*)$. (Množina vrcholů V je sjednocení množin lexikálních jednotek obou slovníků: $V = \{A,B,C,D,E,F\} \cup \{G,H,I,J,K\}$.)



Obrázek 5.2: Lexikální jednotky slovesa *obracet*: vlevo je šest LU uvedených ve VALLEXu, vpravo pět rámců z PDT-Vallexu. Zápis neodpovídá přesně ani jednomu ze slovníků, jedná se o vizualizaci společného formátu SAMR, ve které jsou zachyceny všechny relevantní údaje. Na tomto obrázku jsou zatím znázorněny veškeré možné páry LU, z nichž musíme vybrat ty správné. Na následujících obrázcích 5.3 a 5.4 budeme postupně ukazovat průběh provazování.

Valenční rámec. Hlavním zdrojem informace měly být valenční rámce. Oba slovníky byly, jak už jsme zmínili, tvořeny na syntaktickém základě, delimitace významů do lexikálních jednotek probíhala primárně podle valenčního rámce. Dále se přihlíželo také k dalším ověřitelným údajům (například zda není možné dvě LU odlišit na základě (ne)možnosti tvořit reflexivní variantu). A v neposlední řadě také došlo k rozdělení, pokud se dva jinak shodní kandidáti na jedinou lexikální jednotku výrazně lišili svými významy.[15] Přestože se nakonec ukázalo,

[15] Příkladem může být třeba sloveso *držet*, které má ve VALLEXu tři LU se shodným rámcem ACT(1) PAT(4) BEN(3)$^{\text{typ}}$. Jeden je vyčleněn pro význam *„rezervovat“* (ve slovníku

že více informace přinášejí lemmata použitá v příkladech a v glosách, zůstává informace z valenčního rámce důležitým indikátorem. (Lemmata, viz níže, přinášejí informaci pouze pozitivní; pokud však nejsou nalezena shodná lemmata, neříká to o dvojici LU prakticky nic.) Valenční rámec je navíc v mnoha případech indikátorem jediným – jen ten je přítomen ve všech LU.

Potřebujeme porovnat dvojici valenčních rámců a určit ty dvojice, které vykazují podstatnou podobnost. Pro tento účel používáme sadu pravidel, která si vzápětí lépe ukážeme na příkladech. Pravidla fungují jako šablony: některé prvky rámce vyžadují v přesném tvaru, jiné nechávají nespecifikované. Dvojice LU se mohou podobat a lišit různým způsobem – každá taková varianta podobnosti má své vlastní pravidlo. Toto pravidlo specifikuje co možná nejobecnější rysy páru, který chceme označit jako podobný. Pravidlo má dvě části: jedna specifikuje tvar rámce v jednom slovníku, druhá ve druhém, přičemž druhá část se již může odkazovat na skutečný tvar konkrétního rámce z prvního slovníku, který byl v šabloně podspecifikovaný. Pokud pravidlo například připouští „*nějaký* fakultativní funktor, nebo nic“, můžeme vyžadovat ve druhém slovníku „*tentýž* fakultativní funktor, nebo nic“. Navíc má pravidlo přidělenu ještě číselnou hodnotu mezi nulou a jedničkou, která vyjadřuje důvěryhodnost tohoto pravidla, jeho předpokládanou spolehlivost. Tato hodnota je v případě, že pravidlo zabere, přičtena ke *Score*. Hodnotu důvěryhodnosti pravidla jsme určovali empiricky, obvykle hned při zápisu pravidla.

↪ *Příklad (Pravidlo č. 1):* Rámec ve VALLEXu obsahuje pouze ACT v bezpředložkovém pádě a PAT, který může být fakultativní. Rámec v PDT-Vallexu obsahuje také pouze tyto dva funktory: první funktor má stejnou podmínku (bezpředložkový pád), v případě PAT musí být zachována obligatornost/fakultativnost, ale mohou přibýt nějaké morfematické varianty; nikoli ubýt. (Formální zápis pravidla přineseme níže.) Toto pravidlo má nastavenou středně vysokou důvěryhodnost (0,6), neboť ačkoli popisuje celý rámec (nepovoluje přítomnost dalších

označeno jako idiom *„držet kamarádovi místo“*), druhý pro význam *„platit, vydržovat“* (ve slovníku označeno jako idiom *„držet dětem učitelku“*) a třetí je pro LU *„dodržovat“* (s příklady *„držet někomu smutek / půst / stráž / svátek / slovo“*), což jsou kandidáti na analytické predikáty, které však zatím nejsou ve VALLEXu komplexně zpracované.

Případně sloveso *lemovat* v PDT-Vallexu, které má jen tři LU a všechny ve tvaru ACT(1) PAT(4). První v obecném významu *„stromy lemují řeku“*, druhý v původním specifickém *„švadlena lemuje sukni“* a třetí v přeneseném *„souhlas ze zasedání je lemován ekologickými projekty“*.

Posledním příkladem, který uvedeme, je sloveso *odnášet*, které má rámec pro *„odebírat někomu něco“* ze sémantické třídy transport, například *„zloděj mu odnášel počítač“* a jiný rámec pro *„nesením někomu dopravovat“* ze třídy exchange: *„odnášela psovi zbytky jídla“*.
odnášet$^{V}_{2}$: ACT(1) ADDR(3) PAT(4) INTT(k+3, na+4, inf)^{opt} ↑DIR^{typ}
odnášet$^{V}_{4}$: ACT(1) ADDR(3) PAT(4) INTT(k+3, na+4, inf)^{opt} ↑DIR^{typ}

valenčních pozic), je tento rámec poměrně krátký a bohužel je tedy pravidlo aplikovatelné na větší množství LU s krátkým rámcem.
Toto pravidlo splňuje například dvojice
usilovat$_1^V$ ACT(1) PAT(o+4)
usilovat$_1^P$ ACT(1) PAT(cont, inf, o+4).

↪ *Příklad (Pravidlo č. 5):* Složitější pravidlo může například vyžadovat v obou slovnících ACT v bezpředložkovém pádě, obligatorní, či fakultativní PAT (ale stejně v obou slovnících) s libovolnou, ale stejnou morfematickou realizací. Dále rámec musí obsahovat jeden ze tří zbylých aktantů (tedy ADDR, ORIG, nebo EFF), v obou slovnících stejný a také se shodnou obligatorností. Morfematické formy v PDT-Vallexu musí být podmnožinou těch z VALLEXu. Rámec v PDT-Vallexu už nic jiného obsahovat nesmí, zatímco ve VALLEXu jsou povolené další – ale pouze typické – valenční pozice. Takové pravidlo je mnohem specifičtější a předpokládáme, že pokud se uplatní, označuje odpovídající si rámce s vysokou pravděpodobností, proto má vysokou hodnotu pro důvěryhodnost (0,9).
Toto pravidlo splňuje například dvojice lexikálních jednotek
chránit$_1^V$ ACT(1) PAT(4) EFF(od+2, proti+3, před+7, aby+V, ať+V)$^{\text{opt}}$
BEN(3)$^{\text{typ}}$MEANS(7)$^{\text{typ}}$
chránit$_1^P$ ACT(1) PAT(4) EFF(proti+3, před+7)$^{\text{opt}}$.

Tato slovně zapsaná pravidla je potřeba formalizovat. Přepsali jsme oba rámce do linearizované formy (velmi podobné té, jakou využíváme v této práci, viz níže) a pravidla jsme zapisovali ve formě regulárních výrazů;[16] včetně možnosti uložit část prvního rámce do proměnné a použít ji při porovnávání s druhým rámcem. Pro zjednodušení zápisu jsme také umožnili často používané části regulárních výrazů definovat jako pojmenované proměnné a dále je snadno používat. Poslední důležitá informace je pořadí slovníků: zda se má nejprve porovnat regulární výraz s PDT-Vallexem, uložit části rámce do proměnných a potom druhý regulární výraz porovnat s VALLEXem, nebo opačně.

Zápis valenčního rámce je kompromisem mezi běžně používanými zápisy pro VALLEX a pro PDT-Vallex a vypadá takto: Každá valenční pozice začíná zkratkou funktoru velkými písmeny. Před ním může být znak pro fakultativnost (?), nebo pro typický funktor (+); bez značky je pozice obligatorní. Pokud má funktor uvedené možné morfematické realizace, následuje bezprostředně za zkratkou funktoru kulatá závorka. V ní jsou symbolické zápisy jednotlivých specifikovaných forem oddělené čárkou. Obvykle je forma symbolizovaná číslem vyjadřujícím pád, nebo předložkou, znakem + a číslem pádu. Existují však také značky např. pro infinitiv, vedlejší větu, přímou řeč atd., aby bylo lze zachytit všechny informace

[16] Z mnoha dialektů regulárních výrazů jsme převzali syntax z Perlu, neboť celý systém je naprogramován v tomto jazyce.

z valenčního rámce uvedené ve slovníku. Stranou stojí již zmíněné závislostní zápisy složitějších konstrukcí pocházející z PDT-Vallexu, které ponecháváme v zápise převzatém z Mikulová et al. (2006). Pro porovnání jsme formy v obou slovnících seřadili abecedně. Po vyjmenování všech forem je závorka ukončena a následuje další funktor. I funktory jsou řazeny podle přesných pravidel:

- První jsou aktanty v pořadí ACT, PAT, ADDR, ORIG, EFF,[17] bez ohledu na to, zda jsou obligatorní, či nikoli,
- následují obligatorní volná doplnění seřazená abecedně,
- po nich kvazivalenční doplnění (fakultativní i obligatorní) seřazená abecedně
- a na konci zbylá (tedy typická) volná doplnění, opět v abecedním pořadí.

Nyní bychom se měli zmínit o zápisu typických funktorů, jak jsme slíbili výše. V PDT-Vallexu nejsou typická volná doplnění značena systematicky. Součástí jejich rámce jsou pouze obligatorní volná doplnění. Naštěstí je zde ale často takový typický funktor připsán k příslušnému slovu, pokud se vyskytne v některém z příkladů. Pro sloveso $chtít^{P}_{1}$ tak můžeme číst: *„chce od rodičů stůl; (ono) to chce nevzdat; ch. auto pro manželku.BEN; ch. za výpomoc.CAUS nový byt; SUBS co za to (místo toho) ode mě chceš; EXT[za+4] ch. za deset korun los;“* Při převodu PDT-Vallexu do formátu SAMR jsme se ještě rozhodli tento zápis zachovat, nepřišlo nám vhodné takto významně zasahovat do valenčního rámce, připisovat do něj typická volná doplnění, které z něj autoři záměrně vyloučili, a generovat k nim morfematické formy. Proto to musíme řešit až nyní při linearizaci valenčního rámce, kdy rámec z PDT-Vallexu vždy doplníme všemi funktory nalezenými v příkladech. Problém nastává s morfematickými formami takového funktoru. Ideálně by bylo potřeba převést *„example: politicky se angažoval až letos ve sporu.REG“* v PDT-Vallexu na zápis ve VALLEXu: $\mathsf{REG(v{+}6)^{typ}}$, což kvůli homonymii českých pádů není možné v kratičkém kontextu učinit jednoznačně. Ve výjimečných případech pak ani není funktor přiřazen ke slovu, viz funktor SUBS v příkladu výše. (Naopak explicitně zapsaná forma jako u funktoru EXT, která by pro nás byla nejvhodnější, se v celém PDT-Vallexu vyskytuje přesně dvakrát.) Proto jsme na morfologickou analýzu rezignovali a jako formu uvádíme přesně ono dotčené slovo, což může ve výjimečném případě (např. idiom) skutečně pomoci. Hlavně to ale znamená, že na formu typických valenčních doplnění nemůžeme při porovnávání brát zřetel.

↪ *Příklad:* V následujících dvou odstavcích zajdeme do technických detailů; ukážeme a vysvětlíme jednoduché skutečně používané pravidlo. Pravidlo používá

[17] Toto pořadí je výhodnější než systémové uspořádání aktantů, neboť kvůli principu posouvání nemůže dojít k vynechání druhé pozice (PAT), pokud jsou obsazeny pozice následující. Takové vynechání by všechna pravidla zbytečně komplikovalo.

syntax regulárních výrazů z jazyka Perl; regulární výrazy ani tuto konkrétní syntax zde nebudeme rozebírat a budeme předpokládat, že je čtenáři alespoň částečně známa.

Pravidlo č. 1 pro VALLEX z prvního výše uvedeného příkladu pak vypadá zapsané regulárním výrazem takto:

Zápis pravidla č. 1, část pro VALLEX

```
1 $act = ACT\(.\)
2 $form = [^(),]
3 $formchunk = (?:,$form+)

4 /$act
5        (\??PAT)
6                    \(
7                        ($form+)
8                        ($formchunk*) ($formchunk*)
9                        ($formchunk*) ($formchunk*)
10                       ($formchunk*)
11                   \)
12  $/
```

Řádky 1 až 3 jsou definice proměnných, následuje samotné pravidlo. Jeho přesný význam je: ACT v libovolném pádě (řádek 4, definice na ř. 1), bezprostředně následovaný PATientem (5), před nímž může, ale nemusí být otazník značící fakultativnost. Morfematická forma PATientu (tedy řetězec v závorce za PAT mezi řádkami 6 a 11) sestává z jedné a více[18] variant, které jsou oddělované čárkou. První člen (7) je tedy samostatná jedna morfematická forma definovaná jako cokoli mimo čárku a závorku (viz řádek 2) a zbylé nepovinné členy jsou definované jako čárka a další morfematická forma (3 a 8–10), což se navíc může opakovat. Dolar na konci (12) pak zabraňuje tomu, aby tento výraz nalezl také rámce, které dál pokračují.

Obyčejné závorky (na řádkách 5 a 7–10) slouží k uložení částí valenčního rámce do proměnných. Takto se tedy můžeme v druhém regulárním výrazu, který se bude porovnávat s PDT-Vallexem, odkazovat na PAT kvůli jeho případné fakultativnosti ($1) a na všechny morfematické formy (rozdělené do $2 až maximálně

[18] Tyto varianty jsou rozděleny (až) do šesti skupin, se kterými je poté možné pracovat. Pro počet šest není žádný skutečný důvod. Ale prostředek, který jsme zvolili pro reprezentaci pravidel, nás bohužel v této věci omezuje: v některých případech je nutné vyjmenovat všechny prvky a tím také určit jejich počet. Více v sekci 5.1.7, v ní Omezení daná regulárními výrazy.

$7), které máme v obou slovnících seřazené stejně. Regulární výraz pro PDT-Vallex, který by přesně kopíroval rámec z VALLEXu a vyžadoval by s ním shodu, by tedy vypadal takto:

```
/$act $1 \( $2 $3 $4 $5 $6 $7 \) $/
```

Skutečný výraz, jak jsme ho popsali pro pravidlo č. 1 výše, je jen benevolentnější a mezi každé dvě morfematické formy (a také před první a za poslední) umožňuje vložit libovolný počet dalších forem (zapsaných jako `$nobracket = [^()]`):[19]

Zápis pravidla č. 1, část pro PDT-Vallex

```
/$act $1 \(
                  (?:$nobracket+,)?
               $2 (?:,$nobracket+)? $3 (?:,$nobracket+)?
               $4 (?:,$nobracket+)? $5 (?:,$nobracket+)?
               $6 (?:,$nobracket+)? $7 (?:,$nobracket+)?
         \) $/
```

Regulární výraz opět končí dolarem, takže ani ve druhém slovníku nejsou povoleny žádné další valenční pozice.

Popsali jsme porovnávání dvojice lexikálních jednotek na základě valenčního rámce a všech jeho součástí. To je první a nejkomplexnější ze čtyř hledisek, která k porovnání můžeme použít. Následují lemmata, reciprocita a kontrola.

Množina lemmat. Nyní popíšeme, jak k porovnání lexikálních jednotek používáme lemmata uvedená v hesle. Myšlenka je jednoduchá: prototypická slova, která se obvykle s daným rámcem pojí, se mohou vyskytovat v obou slovnících – a v takovém případě je to velmi významné vodítko. Navíc oba slovníky z historických důvodů sdílejí příklady u některých raných hesel téměř doslovně.

↪ *Příklad:* Pro naše sloveso *obracet* na obrázku 5.3 je to případ fráze *„obracet pozornost“* (uvedené v příkladech v LU C i I) a *„obracet naruby“* (v LU D a H). Nastává to ale též pro doslovný význam (fyzicky nějaký předmět otočit v prostoru), tedy pro lexikální jednotky A a J: *„obracet skříň“*, *„obracet ke zdi“* a *„obracet auto“*, za což pravděpodobně alespoň částečně vděčíme společnému základu obou slovníků.

Jak už jsme zmínili, tento zdroj informace – vedle porovnání valenčních rámců – pomohl více. Poměrně často se pro nějakou dvojici lexikálních jednotek uplat-

[19] Při zpracování pravidla (což je vlastnost regulárních výrazů) se vyzkouší veškerá možná rozdělení morfematických forem ve VALLEXu do šesti skupin a pokud existuje alespoň jediné, při kterém jsou přebytečné formy z PDT-Vallexu vloženy do mezer mezi tyto skupiny, je pravidlo úspěšné.

nil[20] a v takovém případě byl dostatečně spolehlivý, takže jsme mu přiřadili vyšší ohodnocení.

Lemmata získáme lemmatizací všech slov, která se vyskytují

- v příkladových větách,
- v glose (pouze v případě VALLEXu),
- jako morfematické realizace funktoru DPHR („part of compound phraseme“) a
- jako morfematické realizace funktoru CPHR („dependent part of phraseme“) (pouze v případě PDT-Vallexu).

Ze získaných lemmat vytvoříme samostatné množiny, jedna lexikální jednotka jich může mít přiřazenu celou řadu: za každou glosu, za každý jednotlivý příklad a případně ještě za fráze. Nerozlišujeme, zda jde o množinu pocházející z příkladu či z DPHR. Z těchto množin odstraníme veškerá synsémantická lemmata a také slovesné lemma odpovídající právě zpracovávanému lexému. Takto pročištěné množiny autosémantických lemmat porovnáme mezi sebou a hodnota dvou nejpodobnějších z obou slovníků bude hodnotou, která se přičte ke *Score*.

Podobnost dvou množin L1 a L2 počítáme jako aritmetický průměr poměru shodných lemmat ku všem v jedné a druhé LU, tedy

$$\text{podobnost}_{L1,L2} = \frac{\frac{|L1 \cap L2|}{|L1|} + \frac{|L1 \cap L2|}{|L2|}}{2}.$$

Nabývá tedy hodnot mezi nulou a jedničkou.[21]

↪ *Příklad: „Obracet pozornost“*: Lexikální jednotka C nám ze své položky „příklad“ poskytne množinu *„obracet pozornost zájem zřetel“*, lemmatizací se nezmění a vypustíme z ní sloveso *obracet*. Tuto množinu budeme porovnávat s příkladem z LU I *„obracet pozornost jinam“*, odkud také vypustíme *obracet*. Společné lemma je pouze *„pozornost“*, poměr ke zbylým lemmatům je tedy 1:3 a 1:2. Podobnost je jejich aritmetický průměr, vychází tedy $(1/3 + 1/2)/2 = 5/12$. To sice není příliš vysoká hodnota podobnosti, ale nejsou to jediné množiny,

[20] Více než čtvrtině ze všech dvojic, na které se úspěšně aplikovalo nějaké pravidlo, zvýšila následně *Score* ještě lemmata. Nejdůvěryhodnější dvojice byly lemmaty podpořeny téměř vždy (95 % dvojic se *Score* > 1 v něm má obsaženo ohodnocení za lemmata).

[21] Experimentovali jsme s dalšími metrikami, zkoušeli jsme mj. Jaccardův koeficient podobnosti definovaný jako prostý poměr společných prvků ku všem prvkům ve sjednocení:

$$J(L1, L2) = \frac{|L1 \cap L2|}{|L1 \cup L2|}.$$

Neshledali jsme výrazné rozdíly, které by hovořily ve prospěch některé z metrik. Ostatně vzhledem k množství koeficientů, vah a parametrů, se kterými pracujeme, ovlivňuje jejich hodnota systém více než drobné odchylky mezi metrikami.

které pro tyto LU porovnáváme. Při porovnání DPHR *„pozornost“* v C a CPHR *„pozornost“* v I získáme shodné jednoprvkové množiny, takže tato (a zároveň výsledná) podobnost je 1.

↪ *Příklad: „Obracet naruby“*: Z lexikální jednotky D ve VALLEXu získáme *„obracel vše vzhůru nohama / naruby“* a upravíme na *„vzhůru noha naruby“*. Z PDT-Vallexu podobně *„obracet vše naruby“* → *„naruby“* (LU H). Podobnost vychází 0,67: $(1/3 + 1)/2 = 2/3$.

Reciprocita. Třetí informací, která je nějakým způsobem obsažena v obou slovnících a bylo by ji tedy lze použít pro porovnávání, je *reciprocita*.[22] Ve VALLEXu je možnost recipročního vyjádření zachycena explicitně: v atributu `rcp`, kde jsou uvedeny participující valenční doplnění (výčtem funktorů) a obvykle také příkladová věta.

Složitější je to s PDT-Vallexem – tam se přímo v hesle možnost recipročního užití nezaznamenávala. Ovšem díky propojení PDT-Vallexu s daty PDT lze částečně tuto informaci získat. Vyhledali jsme všechna reciproční užití, která se v PDT vyskytují, a funktory participujících členů jsme přiřadili patřičným lexikálním jednotkám v PDT-Vallexu (každé sloveso je instancí nějaké LU a vede od něj na tuto LU do slovníku odkaz), čímž jsme slovník pro naše účely obohatili. Omezení je, že LU můžeme doplnit o reciprocitu pouze tehdy, když byla LU v datech PDT recipročně použita. Je zřejmé, že u málo častých LU a u málo obvyklých, leč možných recipročních užití to je problém.

S PDT-Vallexem rozšířeným o reciproční informaci už jednoduše zjistíme, zda obě jednotky v porovnávané dvojici LU umožňují reciproční užití. Pokud ano a dokonce stejného typu (tedy s užitím stejných funktorů), zvýšíme *Score* o 1 (opět empiricky přiřazené hodnoty). Pokud se shoduje alespoň jeden funktor, může jít o odlišnost danou rozdíly v přiřazování funktorů mezi slovníky a přičítáme ke *Score* 0,7. Konečně 0,5 bodu přičteme, pokud obě LU umožňují reciprocitu, jen jiného typu.

↪ *Příklad:* Lexikální jednotky B i F z obrázku 5.3 jsou obě spojeny s LU G v PDT-Vallexu. Při bližším pohledu na informace uvedené na obrázku zjistíme, že není možné rozhodnout, která z nich (nebo zda obě) odpovídají LU G. Ovšem pouze jedna z nich (B) je doplněna o atribut `rcp`, který upřesňuje, že ACT s PAT mohou být použity recipročně. Pokud by byla LU G použita v PDT recipročně, umožnilo by to výrazně posílit důvěryhodnost spojení B–G.

[22] Reciprocitou (viz sekce 4.2, strana 28) rozumíme možnost symetrického použití dvou a více valenčních doplnění dané lexikální jednotky.

Kontrola. Stejným způsobem by bylo možné použít také informaci o *kontrole*.[23] Kontrola je zaznamenávána jen ve VALLEXu. Do PDT-Vallexu je (podobně jako v případě reciprocity) pouze možné ji doplnit, opět z dat PDT.

Na tomto místě předjímáme výsledky z diskuse (sekce 5.1.7, v ní Reciprocita a kontrola) a rovnou uvedeme, že na základě menšího užitku z reciprocity, než jsme očekávali, a na základě toho, že kontrola se v datech vyskytuje podobně řídce, jsme zde předpokládali malý dopad a nakonec jsme kontrolu vůbec nepoužili.

Score

Zopakujme, že každou dvojici lexikálních jednotek jsme postupně ohodnotili na základě podobnosti jejich valenčních rámců, shody lemmat a schopnosti vstupovat do recipročního vztahu. Součet těchto tří hodnocení tvoří *Score*, což je zároveň ohodnocení hran v grafu, který nazveme $\mathsf{G}' = (\mathsf{V}, \mathsf{E}')$. G' je podgraf úplného bipartitního grafu G^* z obrázku 5.2; V je stejná množina všech LU z obou slovníků, $\mathsf{E}' \subseteq \mathsf{E}^*$ je množina hran, které získaly nenulové *Score*. Čím jsou si dvě jednotky podobnější a čím spíše by měly být ve výsledku propojeny, tím vyšší *Score* by měly dostat. Pro sloveso *obracet* jsou takto získaná *Score* znázorněna na obrázku 5.3. Stanovíme si práh (v našem případě 0,2, což vycházelo při testech na malých datech nejlépe) a pokud ho *Score* nepřekročí, považujeme ho za nulové. Propojené dvojice jsou pak kandidáty na výsledné prolinkování.

Výběr pokrývajících dvojic

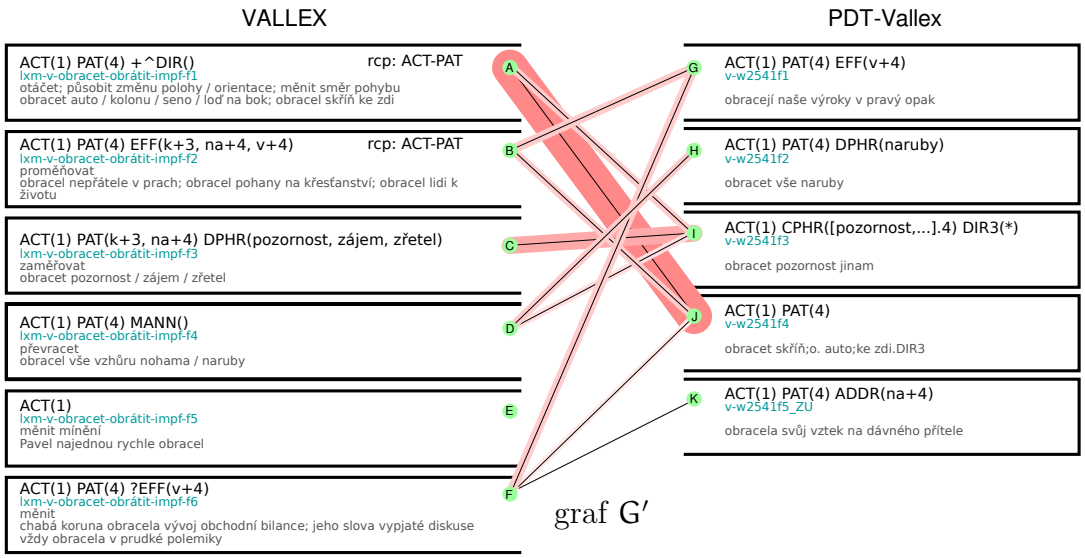
V ideálním případě by ke každé lexikální jednotce byla přiřazena právě jedna jednotka z druhého slovníku (perfektní párování zmíněné v sekci 5.1.3). Je patrné, že obrázek 5.3 tento ideál nesplňuje ani přibližně (například LU F je spojena hned se třemi dalšími jednotkami, zatímco jednotka E není spojena s žádnou). Příčinou je kombinace odlišné delimitace lexikálních jednotek v obou slovnících a chyb v ohodnocení, která nám velmi často neumožní spojit LU jedna ku jedné.

Přesto se budeme snažit omezit počet hran z jednoho vrcholu. Níže popsaným hladovým algoritmem se snažíme dodržet a vyvážit dva principy:

- eliminovat vrcholy stupně vyššího než jedna (blížit se párování) a
- naopak nevytvářet izolované vrcholy (nevzdalovat se hranovému pokrytí).

Chceme vytvořit podgraf $\mathsf{G} = (\mathsf{V}, \mathsf{E})$, kde $\mathsf{E} \subseteq \mathsf{E}'$. Hrany z E', které mají *Score* vyšší než 1, zařadíme do výsledné množiny E všechny. V prvním průchodu (od nejlépe ohodnocených hran) přidáme hranu $(\mathsf{v}_\mathsf{i}, \mathsf{p}_\mathsf{j})$, pokud oba vrcholy v_i i p_j

[23] Pro vysvětlení kontroly viz sekci 4.2 na straně 28.



Obrázek 5.3: Sloveso *obracet* po přiřazení *Score* každé dvojici LU. Vyšší *Score* je reprezentováno tmavší a silnější hranou.
(Pro účely ukázky je použit starší neodladěný výstup s řadou chybných spojení, což využijeme při odstraňování nevhodných hran směřujíce k obrázku 5.4.)

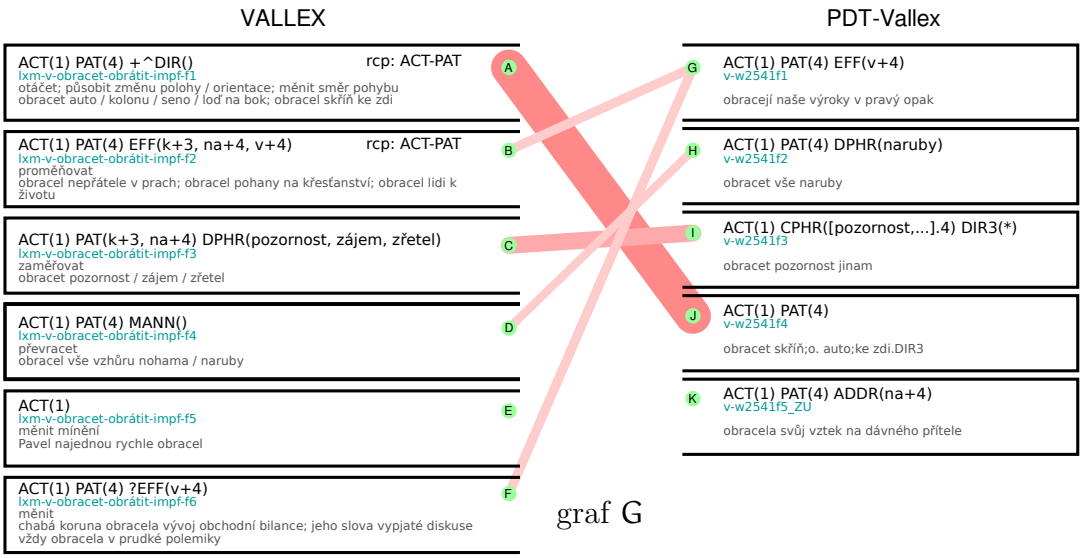
jsou v G izolované. V druhém[24] (opět hladovém) průchodu zbylými hranami již přidáme hranu, pokud je v G izolovaný jeden z vrcholů v_i či p_j.

Tímto postupem z obrázku 5.3 správně eliminujeme například hranu (A, I), neboť z obou vrcholů již vede hrana s vyšším *Score*. Vrchol G bude mít ve výsledku stupeň 2, aby byl pokryt vrchol F (neboť hrana (F, K) nepřekročila stanovený práh). Výsledný graf G s vybranou podmnožinou hran E je na obrázku 5.4.

Ruční kontrola

Pokud bychom chtěli zaručit spolehlivé výsledky za cenu ruční práce (mít tak tedy jen *polo*automatické propojení slovníků), zde přichází prostor pro ruční kontroly a opravy. (My jsme však prozatím ruční práci nevyužili a celý proces ponechali automatický.)

[24] Graf, kvůli kterému potřebujeme dva průchody, vypadá zapsaný maticí sousednosti například takto: $\begin{matrix} \\ c \\ d \end{matrix}\begin{matrix} \begin{matrix} a & b \end{matrix} \\ \begin{pmatrix} 1,5 & 0 \\ 0,8 & 0,7 \end{pmatrix} \end{matrix}$. Chtěli bychom získat výsledný graf s hranami (a, c) a (b, d). Při jediném průchodu umožňujícím spojit izolovaný vrchol s již pokrytým bychom však ve výsledném grafu měli všechny tři hrany.



Obrázek 5.4: Výsledné mapování slovesa *obracet*. Hrany E, které jsou zde zachyceny, byly vybrány jako podmnožina E′ z obrázku 5.3. Příslušné dvojice lexikálních jednotek tedy budou propojeny napříč slovníky (pokud případná ruční kontrola nerozhodne jinak).

Můžeme roztřídit spojené páry LU do tří skupin podle hodnoty dosaženého *Score*. Dvojice v první skupině pak rohlásit za zaručeně správně spojené a dál se jim nevěnovat. Lexikální jednotky v druhé skupině předkládat pár za párem anotátorovi, kterému stačí bez kontextu ostatních částí lexému potvrdit či odmítnout navrženou dvojici, což je relativně rychlé (ale nevyhneme se malému riziku chyb). Třetí skupinu nejméně spolehlivých dvojic nabídnout anotátorovi k revizi a opravám. Zde je potřeba mu zobrazit i všechny ostatní LU v lexému, aby se mohl kvalifikovaně rozhodnout, jak případné chyby opraví, což je sice náročné, ale po automatickém odstranění nejslabších hran v minulém kroku jich nebude moc.

Chtěli bychom podotknout, že pečlivé, časově náročné a drahé ruční anotace[25] nejsou naším cílem, neboť jde o úlohu poměrně náročnou i pro lidské anotátory. V tabulce 5.3 ukážeme, že i v případě slovesných lexémů, které obsahují pouze dvě až tři LU, se dva anotátoři v průměru jednou z deseti případů neshodnou.

[25] Mluvíme pouze o anotaci odpovídajících si LU, bez možnosti oba slovníky editovat, neboť slovníky se – jak jsme již uvedli – dále a souběžně vyvíjí, každý podle svých nastavených pravidel. To pochopitelně nevylučuje umožnit anotátorovi připojit poznámku, na základě které pak bude možné v jednom či obou slovnících nějakou nalezenou chybu opravit.

Nutným krokem po ručních opravách by bylo zaznamenání všech změn pro opakování mapování v budoucnu. Počítáme s tím, že po čase bude nutné nové verze vyvíjejících se slovníků opět propojovat. Po aplikaci automatických pravidel na pozměněné lexikální jednotky bychom opravili známé problémy zaznamenané z minulých ručních oprav.

Nakonec jsme se ovšem vydali cestou automatického propojení, bez ručního zásahu. V závislosti na tom, jak se získaná data osvědčí (jaké bude jejich využití a kolik budou obsahovat chyb), není vyloučeno, že tuto otázku v budoucnu ještě otevřeme a k ručním opravám se vrátíme.

5.1.5 Evaluace

V této sekci přineseme výsledky mapování. Seznámíme čtenáře s ručně anotovanými testovacími daty, proti nimž vyhodnocujeme výsledky, s jednoduchou metodou (baseline) pro srovnání a se způsobem, jakým výpočet úspěšnosti provádíme.

Statistiky

Nejprve se ještě krátce zastavme u velikosti slovníků a objemu dat v nich obsažených, což popisuje tabulka 5.2 (rozšíření tabulky 5.1 ze sekce 5.1.1 Rozsah slovníku).

Čísla z první části „Lexémy“ mají význam zejména pro VALLEX, neboť v PDT-Vallexu lexémy nesdružují vidové protějšky sloves ani jejich varianty. Lexémy, neboli slovníková hesla tak v PDT-Vallexu přímo odpovídají lemmatům. Naproti tomu VALLEX v jednom slovníkovém heslu obsahuje průměrně téměř dvě lemmata (řádek D), některé lexémy pak reprezentují až osm různých lemmat.[26] Řádek B ukazuje, do kolika lexikálních jednotek byl delimitován význam celých lexémů. Jejich počet je tedy mnohem nižší než celkový počet LU po rozgenerování lexémů pro jednotlivé vidové protějšky a varianty sloves (řádek I).[27] V této části jsou také v PDT-Vallexu uváděny počty lexikálních jednotek pro

[26] Příkladem je lexém obsahující tři varianty nedokonavého vidu: *vystřihávat, vystříhávat-I, vystřihovat*, po dvou variantách pro obě možná vyjádření dokonavého vidu: *vystřihnout, vystříhnout* a *vystřihat, vystříhat-I* a navíc ještě iterativum *vystřihovávat*. Pokud by se těchto osm lemmat vyskytovalo v PDT-Vallexu, byly by reprezentovány osmi různými (ačkoli podobnými) slovníkovými hesly, osmi lexémy. (Stejně tomu je po převodu VALLEXu do formátu SAMR.)

[27] Když počet LU v lexémech (řádek B) vynásobíme **průměrným** počtem lemmat v lexému (D), získáme hrubou představu o celkovém počtu lexikálních jednotek: $6\,451 \times 1,79 = 11\,547$. Protože ne každá lexikální jednotka se vztahuje ke všem lemmatům sdruženým v lexému, je skutečný počet uvedený v třetí části tabulky (11 229 v řádku I) lehce nižší.

			VALLEX ver. 2.6	PDT-Vallex ver. 2.0
Lexémy	A	Počet lexémů ve slovníku	2 726	11 656
	B	Počet LU v lexémech	6 451	17 720
	C	Průměrný počet LU na lexém (=B/A)	2,36	1,52
	D	Průměrný počet lemmat na lexém (=E/A)	1,79	1
Lemmata	E	Počet lemmat ve slovníku	4 887	11 656
	F	Počet slovesných lemmat ve slovníku	4 887	7 103
	G	Počet různých slovesných lemmat	4 787	7 103
	H	Počet společných lemmat ve slovnících[28]	3 543	3 466
Slovesné LU	I	Počet slovesných LU ve slovníku	11 229	11 933
	J	Počet slovesných lemmat s jedinou LU	2 141	4 931
	K	Průměrný počet LU na lemma (=I/G)	2,30	1,68
	L	Nejvyšší počet LU na lemma	26	138
Společné LU	M	Počet společných lemmat (=H)	3 543	3 466
	N	Počet LU společných lemmat	8 200	7 682
	O	Počet společných lemmat s jedinou LU	1 241	1 815
	P	Průměrný počet LU (=N/M)	2,53	2,22
	Q	Nejvyšší počet LU na společné lemma	26	138

Tabulka 5.2: Počet lexémů, lemmat a lexikálních jednotek zvlášť ve VALLEXu a PDT-Vallexu i v jejich „společné“ části.

všechny slovní druhy. (Můžeme si všimnout, že díky substantivům typicky s jedinou LU je zde poměr LU na slovníkové heslo (1,52) nižší než v třetí části tabulky (1,68) v řádku K, kde už se číslo týká pouze sloves.)

Druhá část „Lemmata“ vypovídá o lemmatech ve slovníku. Připomínáme, že jeden lexém ve VALLEXu může sdružovat několik lemmat, proto když mluvíme o lemmatech (například kdekoli v této tabulce), mluvíme vlastně o expandované verzi slovníku, kterou představujeme v sekci 4.13. VALLEX ve svých lexémech obsahuje téměř 4 900 lemmat (řádek E). PDT-Vallex pak více než dvojnásobek, ovšem velkou část tvoří substantiva, adjektiva a adverbia. Třetí řádek G zohledňuje fakt, že ve VALLEXu jsou homografy zachyceny odděleně (viz poznámka 8 v sekci 5.1.2), odlišuje se tedy například *zapírat-I* (= podpírat), *zapírat-II* (= popírat) a *zapírat-III* (~ zaprat). Proto se třetí řádek liší od druhého pro VALLEX, ale neliší se pro PDT-Vallex, kde by všechny lexikální jednotky pro tato tři

slovesa byly spojeny v jediném hesle *zapírat*^{P}. Poslední řádek H uvádí jen taková lemmata, která se vyskytují v obou slovnících, a týká se jich tudíž propojování.[28]

Třetí část „Slovesné LU“ se týká lexikálních jednotek v celé slovesné části slovníku, čtvrtá část „Společné LU“ je pak její obdobou pro tu část jednotek, která náleží lemmatům zastoupeným v obou slovnících. Řádek I, resp. N uvádí počet lexikálních jednotek slovníku; pro společná lemmata to je počet prvků, které vstupují do naší párovací úlohy. Následující dva řádky pak vypovídají o obtížnosti úlohy: u lemmat s pouze jedinou LU (v obou slovnících!) můžeme předpokládat vysokou úspěšnost (J a O), nicméně mnoho dalších lemmat obsahuje větší množství LU, mezi nimiž musíme vybrat ty správné do párů.

Testovací data

Ruční anotací jsme vytvořili testovací data v rozsahu 200 sloves.

Zavedeme značení pro množinu hesel ze slovníku, která obsahují právě i lexikálních jednotek, $L_i = \{h \in \text{VALLEX}; |h| = i\}$.[29]

Prvních 90 sloves (set_1) bylo vybráno tak, aby rovnoměrně zastupovalo různě komplexní lexémy. Omezili jsme se na lexémy s jedním až devíti lexikálními jednotkami. Do těchto devadesáti sloves se z každé množiny $L_1, ..., L_9$ dostalo náhodně vybraných deset sloves.

Dalších 110 sloves (set_2) bylo vybráno tak, aby proporcionálně odpovídalo mohutnostem množin $L_i, i = 1, \ldots, 26$, tedy všech, které jsou ve VALLEXu zastoupeny (viz řádek Q v tabulce 5.2). Poměr $|L_i|$ k velikosti celého VALLEXu je tudíž obdobný poměru počtu anotovaných sloves s i rámci ($A_i = set_2 \cap L_i$) ku všem 110 vybraným slovesům: $\frac{|L_i|}{|\text{VALLEX}|=3240} \approx \frac{|A_i|}{|set_2|=110}$.

Pro tato slovesa pak dva anotátoři nezávisle na sobě spárovali lexikální jednotky z obou slovníků. Jejich vzájemnou shodu zachycuje tabulka 5.3. Tím jsme získali dvě sady dat, proti kterým můžeme porovnávat naši automatickou metodu i baseline.

Baseline

Za tzv. „baseline“ pro jedno slovesné lemma, s níž se můžeme porovnávat, stanovíme následující metodu. K určenému lemmatu z VALLEXu najdeme shodné

[28] Pokud je lemma z PDT-Vallexu ve VALLEXu přítomno dvakrát – jako dva homografy – a bude tudíž probíhat dvojí porovnávání (lemma proti lemma-I a poté proti lemma-II), počítáme ho jako jedno v PDT-Vallexu a dvě společná ve VALLEXu. Proto se čísla v řádku H mírně liší.

[29] Samozřejmě dané sloveso nemá v obou slovnících stejný počet LU, proto jsme arbitrárně zvolili jeden z nich, VALLEX, a řídili se počtem LU v něm.

lemma v PDT-Vallexu. V obou slovnících seřadíme lexikální jednotky (podle čísel v jejich identifikátorech) a propojíme je v tomto pořadí (první s první, druhá s druhou, …). Žádné LU nepřiřadíme dvě a více ekvivalentů z druhého slovníku. Pokud je v jednom slovníku více LU než ve druhém, zůstanou nepropojené.[30] (V případě homograf přistupujeme k oběma stejně: nejprve popsaným způsobem spárujeme lexikální jednotky s prvním z nich, poté tytéž jednotky spárujeme s druhým.)

Popsaný způsob není tak naivní, jak se může zdát, neboť z pozorování se zdá, že lexikální jednotky byly do slovníkového hesla v obou slovnících vkládány v podobném pořadí (tedy od nejobvyklejších, prototypických užití daného lemmatu k méně častým); případně byla dodatečně zhruba do tohoto pořadí uspořádána.

Způsob vyhodnocení

Pro lepší představu o náročnosti anotátorské úlohy uveďme, že mezi těmito konkrétními vybranými 200 slovesy může existovat až 2 721 párů lexikálních jednotek,[31] z nichž je nutné vybrat ty skutečné dvojice, které si odpovídají. Anotátor A jich takto vybral 529, anotátor B označil za správné 493 dvojic, což představuje méně než pětinu hran. Je samozřejmé, že většina dvojic (jak těch možných, tak anotovaných) se vyskytuje v rámci sloves s vysokým počtem LU (například vybraným deseti slovesům o devíti rámcích náleží 700 potenciálních a 70 správných dvojic), zatímco v celém slovníku má téměř polovina sloves jen jedinou LU.

Všechny výsledky budeme porovnávat pomocí *„precision“* a *„recall“*[32] (ať už výsledky automatické procedury vůči anotovaným datům, nebo anotátory vůči sobě). Vzhledem k uvedené nevyrovnané obtížnosti mezi skupinami L_i počítáme precision i recall zvlášť po jednotlivých skupinách. Výsledná precision je váženým průměrem dílčích výsledků precision $P(L_i)$, které dostanou váhu podle mohutnosti množiny L_i: největší váhu dostane $P(L_1)$, neboť L_1 obsahuje mnohem více sloves než zbylé $L_2, \ldots, L_{26}$. Slovesa s devíti LU sice obsahují více dvojic lexikálních jednotek, nás však zajímá, kolik celých sloves se nám podaří dobře propojit.[33] Recall počítáme obdobně.

[30] Obrázek 5.4 by tedy pro baseline obsahoval tyto hrany: A–G, B–H, C–I, D–J a E–K. Poslední LU F z VALLEXu by zůstala nespojená.

[31] Tedy 2 721 hran ve 200 úplných bipartitních grafech.

[32] Precision, česky někdy též *přesnost* je definovaná jako poměr správně nalezených ku všem nalezeným: $P = \frac{\text{správné}}{\text{nalezené}}$. Recall, česky někdy též *úplnost* je definovaná jako poměr správně nalezených ku všem hledaným: $R = \frac{\text{správné}}{\text{hledané}}$. Chybně nalezené případy („false positives“)

Slovesa z množiny	Počet sloves	Precision anotátora B vůči anotátoru A	Recall anotátora B vůči anotátoru A[34]
A_1	52	100	98
A_2	42	89	94
A_3	27	89	81
A_4	18	85	74
A_5	14	86	76
A_6	12	79	74
A_7	12	91	83
A_8	11	78	73
A_9	10	90	82
(A_{10}	1	73	73)
(A_{13}	1	100	50)[35]
Vážený průměr	200	**93**	**91**

Tabulka 5.3: Anotátorská shoda na ručním mapování 200 vybraných sloves mezi VALLEXem a PDT-Vallexem. Množina A_i obsahuje anotovaná slovesa, která mají ve VALLEXu i LU.

Podívejme se, jak si stojí oba anotátoři při porovnání mezi sebou (tabulka 5.3). Mezianotátorská shoda přes devadesát procent (poslední řádek) ukazuje, že se anotátoři dokážou do značné míry shodnout a úloha tedy má smysl. Přestože se výběr první dávky sloves od druhé významně lišil, popsané vážené průměrování právě tento rozdíl maže, takže výsledky pro set_1 i set_2 vycházejí podobně a v tabulce 5.3 tedy dávky neodlišujeme a uvádíme pouze celková čísla pro každou ze skupin L_i a vážený průměr pro všech 200 sloves dohromady.

Tabulka 5.4 zobrazuje první čísla z automatické procedury. Ukazuje, kolik lexikálních jednotek se nachází ve vybraných 200 slovesech, kolik dvojic spojili anotátoři a kolik bylo spojeno automaticky. Zvýrazněná čísla poukazují na dvě

tedy snižují precision, nenalezené případy („false negatives“) snižují recall. Dále používáme F_1 score, které je harmonickým průměrem obou měr: $F = 2\frac{PR}{P+R}$.

[33] Sloveso, které má řekněme obě dvě své LU správně propojené s druhým slovníkem, je pro nás jistě cennější, než třeba i tři správné dvojice z dvanácti u komplexního slovesa.

[34] …což je pochopitelně totéž jako precision A vůči B.

[35] Slovesa s deseti a se třinácti LU se objevila jen v druhé dávce 110 sloves. Jejich zastoupení jediným reprezentantem odráží nízkou četnost ve slovníku, snižuje ovšem vypovídací hodnotu anotátorské shody na těchto řádcích, proto jsou zde uvedeny v závorce.

		VALLEX	PDT-Vallex
M	Počet anotovaných lemmat	200	200
N	Počet LU anotovaných lemmat	716	528
P	Průměrný počet LU (=N/M)	3,6	2,6

		VALLEX				PDT-Vallex			
Namapované LU		A	B	auto	base	A	B	auto	base
(0 hran)	Vůbec	249	280	**174**	245	61	72	92	61
(1 hrana)	Jednoznačně	415	386	464	471	417	422	312	463
(>1 hrana)	Víceznačně	52	50	78	0	50	34	**124**	4
Počet dvojic LU		529	493	649	471	529	493	649	471

Tabulka 5.4: Počty lexikálních jednotek ve vybraných 200 slovesech. (První sloupec odkazuje na odpovídající řádky tabulky 5.2. Je dobře patrné, že výběrem deseti sloves i z vyšších množin $L_4, \ldots, L_9$ jsme výrazně zvýšili počet LU připadajících na jedno lemma.) Uvádíme výsledky jak pro ruční anotaci (anotátoři A a B), tak pro automatickou proceduru (auto) a pro baseline (base). Nenamapovaná lexikální jednotka („vůbec“ v tabulce) je taková, ze které ve výsledku nevede ani jedna hrana, z jednoznačně namapované vede právě jedna hrana a z víceznačně namapované vede více hran.
Čtyři víceznačné hrany pro baseline vznikly kvůli homografům: první jednotka byla spojena s první lexikální jednotkou postupně pro dva různé lexémy z VALLEXu.

hlavní odchylky automatické procedury od testovacích dat, obě stejným směrem. Abychom ve statistice dopadli podobně, měli jsme ve VALLEXu ponechat více LU nespárovaných (nesnažit se k nim nalézt ekvivalent – podle anotátorů zjevně takový ekvivalent neexistuje častěji, než jsme očekávali) a v PDT-Vallexu jsme měli důrazněji zamezit spojování jedné LU s více „ekvivalenty“ z VALLEXu (mezi nimiž mohou snadno být právě ony nespárované LU). Obojí znamená totéž: Generujeme mnoho zbytečných hran. Měli bychom být „přísnější“ při výběru pokrývajících hran v poslední fázi algoritmu: buď nastavit vyšší práh, nebo vůbec nepřidávat hranu jen proto, abychom se zbavili izolovaného vrcholu – každopádně ponechat méně hran.

Výsledky

Tím se dostáváme k vyhodnocení automatické procedury na vzorku dat vůči manuální anotaci[36] a porovnání s baseline.

V tabulce 5.5 jsou opět jednotlivé řádky počítány odděleně pro jednotlivé množiny $A_1, \ldots, A_9$ (v nichž je alespoň deset anotovaných sloves). Počty sloves v množinách jsou stejné jako v tabulce 5.3. V levé části jsou naše výsledky, vpravo je pro srovnání baseline.

Slovesa z množiny	Automatické mapování			Baseline		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score
A_1	93	76	84	91	89	90
A_2	77	73	75	67	61	63
A_3	63	80	70	37	34	35
A_4	60	74	66	37	34	35
A_5	59	89	71	27	27	27
A_6	51	82	63	23	21	22
A_7	46	69	55	20	19	20
A_8	40	68	50	20	21	20
A_9	51	73	60	22	21	21
Vážený průměr	**78**	**76**	**77**	**65**	**62**	**63**

Tabulka 5.5: Úspěšnost spojování VALLEXu a PDT-Vallexu pro 200 sloves. V tabulce jsou vedle sebe ke srovnání jak naše výsledky, tak i výsledky, jakých dosáhla jednoduchá baseline. Překonali jsme ji o čtrnáct procentních bodů (v F_1 score).

S rostoucím počtem LU pro lemma úspěšnost očekávatelně a viditelně klesá, ačkoli kvůli malému vzorku dat není průběh monotónní. Nejhorší je tudíž úspěšnost u komplexních lexémů (zde množiny A_7 a A_8), kde ale také nejzřetelněji překonáváme baseline.

[36] Automatickou proceduru i baseline porovnáváme proti anotátoru B. Jejich anotace však byly poměrně podobné a čísla počítaná proti anotátoru A jsou téměř stejná.

5.1.6 Výstup z projektu

Předchozí sekce 5.1.5 Evaluace se omezovala na vybraných 200 sloves, neboť k nim existují testovací data a je možné je vyhodnocovat. Teď se vrátíme k celým slovníkům.

Připravenou a otestovanou metodu popisovanou v celé této kapitole jsme spustili na celé slovníky VALLEX a PDT-Vallex. Předpokládáme, že i zde dosahují výsledky zhruba sedmasedmdesátiprocentního F_1 score, stejně jako dosahovaly na náhodném vzorku dat. Výsledek mapování v počtech propojených LU je vidět v horní části tabulky 5.6.

Protože však nechceme zveřejnit chybné odkazy na rámce, které by odkazující LU neodpovídaly, změřili jsme také úspěšnost výsledků upravených tak, že jsme ponechali pouze takové páry, jejichž *Score* přesáhlo hodnotu 1. Výrazná redukce přijatelných párů (téměř na čtvrtinu) zajistila celkovou váženou precision 91 %, což je zcela na úrovni úspěšnosti lidských anotátorů (91 a 93 %). Cenou byl samozřejmě velký propad přiřazených rámců (recall klesl na 27 %). Pouze takto redukované páry jsme se prozatím rozhodli zveřejnit, nová čísla jsou tedy vidět v dolní části tabulky 5.6.

Počet LU	VALLEX	PDT-Vallex	P	R
nenamapovaných (0 hran)	2 245	1 622		
jednoznačně namapovaných (1 hrana)	5 537	4 670	78	76
víceznačně namapovaných (> 1 hrana)	1 034	1 382		
zveřejněných s 1 hranou	1 816	1 795	91	27
zveřejněných s >1 hranou	131	144		

Tabulka 5.6: Výsledky spojování slovníků VALLEX a PDT-Vallex. V pravé části je precision a recall pro původní i zveřejněné výsledky.

Tyto odkazy jsme zanesli jak do datového formátu VALLEXu v XML (ve formě identifikátorů PDT-Vallexu), tak také do prezentační formy na webu (ve formě hypertextových odkazů na web PDT-Vallexu, viz obrázek 5.5). Jedná se o odkazy jak na úrovni lexému (kde pro jednotlivá lemmata obsažená v lexému VALLEXu odkazujeme na odpovídající slovníková hesla PDT-Vallexu), tak zejména na úrovni lexikálních jednotek (LU z VALLEXu obsahuje odkazy na rámce PDT-Vallexu). O tyto odkazy jsme obohatili všechna lemmata, která se vyskytují v obou slovnících, což představuje odkaz do PDT-Vallexu z 88 % lexémů. Aktuální verze VALLEXu (a pochopitelně také plánovaná letošní verze 3.0)

tyto odkazy obsahuje. Podobně příští vydání PDT-Vallexu je již plánováno s odkazy do VALLEXu.

Obrázek 5.5: Ukázka provázání sloves *mít se, mívat se* na webu VALLEXu. V pravém horním rohu je odkaz na celé slovníkové heslo *mít se* se všemi valenčními rámci na webu PDT-Vallexu; odkaz na *mívat se* zde není, neboť toto lemma PDT-Vallex neobsahuje. První a třetí LU (zdůrazněno tečkou vlevo) zde obsahuje odkaz do PDT-Vallexu (pouze však pro lemma *mít se*) a také drobné číslo v závorce, které představuje výsledek *Score* při mapování.

5.1.7 Diskuse

Porovnání dvou paralelních manuálních anotací ukázalo, že úloha je rozumně řešitelná. Nicméně porovnání slovesného hesla ve dvou slovnících a propojení LU je podle naší zkušenosti i podle vyjádření anotátorů často nelehký úkol a rozhodnout, který ze dvou či více přijatelných kandidátů vybrat (nebo zda všechny), je (nejen časově) náročné. Jako u většiny anotací ani zde pochopitelně neexistuje jediné správné řešení, ani zde si anotátor není vždy zcela jist svou konečnou volbou – ani v případě, že dojde ke shodě.

Náročnost úlohy potvrzují i dosažené výsledky. Na jednodušších heslech s menším počtem LU dosahujeme pochopitelně vyšší úspěšnosti než na heslech komplexních. Díky faktu, že těch jednodušších je ve slovníku více než komplexních, dosahujeme uspokojivého průměrného F_1 score 77 %.

Nejprve vyhodnotíme přínos využití informací o reciprocitě a kontrole, poté postupně projdeme pět identifikovaných příčin chybných výsledků a zhodnotíme nevýhody zvoleného porovnávání rámců založeného na regulárních výrazech.

Reciprocita a kontrola

Na straně 104 jsme popsali, jak jsme do PDT-Vallexu doplnili informaci o reciprocitě a využili ji při porovnávání obou slovníků.

Ukázalo se ovšem, že reciprocita je v PDT, odkud jsme informaci o ní získali, natolik řídce zastoupena, že na výslednou úspěšnost nemá valného efektu. V celém PDT 2.0 je 493 výskytů slovesné reciprocity,[37] v nichž se opakuje 144 slovesných rámců pro 130 lemmat. Takové množství pokrývá pouze 1,2 % všech slovesných rámců PDT-Vallex 2.0. Z toho je zřejmé, že nelze očekávat významný dopad na výsledek.[38]

Podobné to je i s další informací, kterou jsme navrhovali doplnit do PDT-Vallexu, a to s kontrolou. Ta se vyskytuje v PDT na 4 706 místech,[39] celkem pro 488 různých rámců (patřících 426 slovesným lemmatům). To je mírně vyšší počet, představuje 4,1 % rámců v PDT-Vallexu. Ovšem zběžná lingvistická analýza naznačila, že se v informaci o typu kontroly neskrývá velký potenciál na vylepšení systému. Fakt, že se jedná o sloveso kontroly, je zakódovaný již v možnosti jedné valenční pozice být vyjádřena infinitivem. Pokud navíc do PDT-Vallexu doplníme informaci o typu této kontroly, tedy o funktoru kontrolujícího členu, nezískáme o mnoho více, protože má-li vůbec jedno slovesné lemma více rámců umožňujících kontrolu, její typ je obvykle shodný. Jak jsme už uvedli dříve, explicitní informaci o kontrole (jinou než infinitivní formu valenční pozice) jsme tedy nakonec nepoužili.

Analýza chyb

Nyní bychom rádi pojmenovali hlavní zdroje chyb ve výstupu naší automatické procedury.

- **Nedostatek informací u jednotlivých LU.** Přes bohatou informaci oba slovníky obsahují nakonec málo záchytných bodů, kterých by se dalo automaticky využít k bezpečnému spojení odpovídajících si valenčních rámců

[37] Tedy 493 t-uzlů s `t_lemma`tem #Rcp, které jsou efektivními potomky slovesného t-uzlu.

[38] Konkrétně pro 90 sloves z množiny set_1 je účinek reciprocity zcela zanedbatelný, neboť se při porovnávání rámců projevila pouze třikrát. Ve srovnání s experimentem bez ní jednou výsledek zlepšila, jednou zhoršila a jednou správný výsledek nezměnila.

[39] Tedy 4 706 slovesných t-uzlů (sloveso kontroly), které mají efektivního potomka (infinitiv), jehož efektivní potomek (kontrolovaný člen) má `t_lemma` #Cor.

(neboť někdy je jí málo i k bezpečnému odlišení různých LU v jednom slovníku). Jak jsme ovšem viděli v případě reciprocity a kontroly, použitelné jsou jen ty informace, které jsou dostatečně četné.

- **Různá delimitace významu.** Dalším problémem, tentokrát mnohem obecnějším, je delimitace významu slovesa do jednotlivých jednotek. Často jsou si jednotlivé lexikální jednotky VALLEXu či valenční rámce PDT-Vallexu natolik podobné, že je i pro člověka obtížné zformulovat, čím se přesně liší (případně proč nejsou sloučeny). Potom je pochopitelně pro člověka (a tím spíše pro automatickou proceduru) velmi náročné k takovým vzájemně podobným jednotkám nalézt odpovídající jednotky ve druhém slovníku, zvlášť pokud i v něm jsou vztahy mezi jednotkami podobně nejasné, či jsou významy zcela jinak rozděleny (například když má sloveso jinou granularitu delimitace významu).

  Toto považujeme za největší problém celého úkolu. Různá granularita a odlišná delimitace významů způsobuje, že hledané propojení nemá být 1:1 a komplikuje celou úlohu. Pokud by se někdo na základě podobnosti slovníků VALLEX a PDT-Vallex (teoretické i strukturní) domníval, že popisovaná úloha je snadná, je potřeba zdůraznit, že slovníky se ovšem liší množstvím a delimitací svých významů. Zkusme si představit dva výkladové slovníky stejného jazyka: jistě budou „vypadat“ podobně, nelze však očekávat, že by snad bylo snadné nalézt odpovídající si významy jednotlivých slov.
- **Chybějící valenční rámce.** Už mnohokrát jsme napsali, že ani jeden z obou slovníků nepokrývá plně všechny existující valenční rámce. V případě PDT-Vallexu chybí rámec jednoduše tehdy, pokud se nevyskytl v PDT, ve VALLEXu sice byla snaha o úplnost, ne však již u idiomů, kterých řada chybí. Valenční rámce, kterým v druhém slovníku chybí protějšek, mají ovšem v kombinaci s předchozím bodem tendenci se s nějakým rámcem přesto spojit, a to za cenu spojení 2:1 apod. Náš systém, který musí hledat nejpodobnější protějšky pro všechny rámce, hledá nejbližšího kandidáta i pro rámec, pro nějž protějšek neexistuje. Vyřešit tento problém lze jen vhodným nastavením prahů, které určí, zda jsou ty nejbližší rámce dostatečně blízké, či nikoli.
- **Odlišné zachycení informací v obou slovnících.** Pokud se vrátíme k obrázku 5.4, vidíme, že stejné valenční rámce jsou v obou slovnících zapisovány různě: mohou se lišit počtem valenčních doplnění (dvojice A–J), jejich typem, tedy funktorem (dvojice D–H), či jejich přípustnými morfematickými formami (dvojice B–G).
- **Lepší pravidla.** Důvody pro vznik chyb jsou samozřejmě i v samotných pravidlech, která používáme, v nastavení jejich důvěryhodnosti, ve vyladění prahů. Dalším ručním porovnáváním rámců a chyb při jejich propojování

bychom mohli toto vše stále vylepšovat. Protože ale nelze tímto způsobem dosáhnout ideálního stavu, jak naznačuje následující příklad, je nutné psaní pravidel někdy ukončit.

↪ *Příklad:* Ve VALLEXu máme následující rámec slovesa (rámce zarovnáváme pro pohodlnější porovnání) *říkat* pro význam žalovat („říkal na něj, že lže“):

ACT(1) PAT(na+4) ADDR(3) EFF(4, cont, že+V)

a v PDT-Vallexu se vyskytuje jen jediný rámec s předložkou *na*:

ACT(1) PAT(na+4)$^{\text{opt}}$ ADDR(3) EFF(4, SPEECH, cont, jak+V, zda+V, že+V)

Tento rámec je ovšem jiný, neznamená žalovat („říkal mu na jeho dotaz podrobně, co dělá“). Správně by se měl spojit s jiným existujícím rámcem VALLEXu, který má tento význam a vypadá následovně (od prvního se liší jen spojkou *zda*):

ACT(1) PAT(na+4) ADDR(3) EFF(4, cont, zda+V, že+V)

V současném stavu spojí naše metoda ukázaný rámec z PDT-Vallexu s oběma podobnými rámci VALLEXu.

Rádi bychom se ještě zastavili u chyb nalezených díky porovnání s ruční anotací. Z nich bychom jistě mohli vyvodit nějaké úpravy pravidel i systému (například tvrdší kritéria pro výběr výsledných hran z kandidátů, jak jsme zmínili v sekci 5.1.5 Evaluace). Ručně anotovaných testovacích dat však máme příliš málo na to, abychom si mohli dovolit je rozdělit na „development test data“, tedy data, na nichž bychom pravidelně testovali experimenty a podle výsledků opakovaně upravovali systém, a „evaluation test data“, na nichž bychom posléze změřili skutečnou úspěšnost upraveného systému. Proto jsme výsledky zveřejnili tak, jak vyšly bez dalších úprav systému.

Omezení daná regulárními výrazy

Regulární výrazy jako základ pro porovnávání valenčních rámců byly jednoduchým a funkčním řešením. Regulární výrazy jsou standardní a ověřený výkonný prostředek a využili jsme jejich rychlé zpracování v Perlu. (Dva příklady jsme uvedli na stranách 98–102 v sekci 5.1.4.) Regulární výrazy při porovnávání navíc efektivně využívají tzv. backtracking, což často využíváme při „uzávorkování“ části výrazu a jeho následném vyhledání: pokud první možný výraz není nalezen, porovnávací proces se vrátí zpět a zkusí „uzávorkovat“ jiný možný řetězec atd. Přestože se v průběhu práce objevilo i mnoho nedostatků a omezení (viz následující bodový seznam), která jsme nepředvídali a která vyžadovala zvláštní pozornost, celkově považujeme zvolené řešení za správné a výhodné.

Předpokládáme, že čtenář je s fungováním regulárních výrazů seznámen. Ze zmíněné sekce jen připomeneme, že úspěšný test vypadá tak, že první regulární

výraz souhlasí s prvním valenčním rámcem a druhý regulární výraz (který může využívat proměnné získané „uzávorkováním“ z prvního porovnání) pak odpovídá valenčnímu rámci z druhého slovníku. V praxi však oba valenční rámce spojíme do jednoho řetězce a také oba regulární výrazy spojíme do jediného výrazu (s použitím stejného oddělovače uprostřed). Test potom provádíme jen jediný s oběma rámci najednou, což nám umožňuje plně využít zmíněný backtracking regulárních výrazů (vysvětlení níže). Jaká tedy byla jejich omezení? Čeho nejde pravidly tvořenými regulárními výrazy (pohodlně, nebo vůbec) dosáhnout?

- **Negace:** Regulárním výrazem nelze říci „cokoli kromě X“. Například pro „kterýkoli funktor vyjma PAT“ jsme definovali proměnnou, která složitým způsobem (se znalostí zbylých funktorů) opisuje potřebnou negaci. Řešení sice není elegantní, nicméně funkční ano.
- **Kombinace prvků v seznamech:** Potřebujeme-li porovnat dva seznamy prvků (například řadu funktorů nebo výčet morfematických realizací jednoho funktoru) a nevyžadujeme-li identitu (chceme pouze nějak specifikovanou podobnost), nejsou pro to regulární výrazy příliš vhodné. Uveďme zde několik případů, kdy je toto porovnání potřeba:
  - Řekněme, že chceme v jednom rámci nalézt (a kladně ohodnotit) „podmnožinu povolených morfematických realizací“ z rámce druhého (jako na straně 98 v sekci 5.1.4). Potom je jednodušší otázku otočit a hledat „nadmnožinu“ realizací z rámce prvního, tedy umožnit vložit mezi každé dvě morfematické realizace nějaké další. Jsme ovšem omezeni tím, že si musíme jednotlivé části v regulárním výrazu „uzávorkovat“, abychom pak mohli mezi tyto označené úseky vkládat ony nadbytečné realizace. Tím ovšem ve chvíli, kdy si je jednotlivě označíme, také pevně určíme, kolik maximálně jich naše pravidlo dokáže pojmout.
  - Kdybychom navíc chtěli porovnat seznam (funktorů, morfematických realizací) proti seznamu s jiným pořadím, nutilo by nás to ke komplikovanému výčtu všech permutací. To jsme však vyřešili už na začátku, kdy jsme obojí setřídili podle stejného klíče.
  - Podobné by to bylo, kdybychom valenční rámce a regulární výrazy pro oba slovníky nespojovali do jediného, jak jsme popsali výše. Obtížné by potom bylo hledání „alespoň jednoho typického funktoru“ (z možného většího množství) v obou valenčních rámcích. Pokud bychom „typický funktor“ vyhodnocovali odděleně, do proměnné `$1` by se uložil jakýkoli první typický funktor v prvním slovníku. Pokud ten by se ve druhém slovníku nenašel, už by se nikdy nevyzkoušely další typické funktory. Právě proto spojujeme oba valenční rámce do jednoho řetězce a porovnáváme ho se spojenými regulárními výrazy: díky tomu se vyhodnocování vrátí zpět a zkusí porovnání s druhým typickým

funktorem. Na tento případ jsou regulární výrazy připraveny dobře. (Stejně jednoduché je tak pravidlo na „dva shodné typické funktory“ apod.)

– Shrňme to: Jsou případy porovnání dvou seznamů, na něž jsou regulární výrazy přímo určené (nalezení jednoho/dvou/tří prvků z více), jiné případy jsme snadno vyřešili předzpracováním (porovnání nesetříděných seznamů), ale některé případy nás nutí k omezením, která bychom jinak nezaváděli (nalezení podmnožiny seznamu).

- **Drobné rozdíly:** Pokud chceme povolit, aby se rámce byť i nepatrně lišily (například je vynechaný jeden funktor, nebo v libovolném doplnění na libovolném místě chybí morfematická varianta), je nutné přesně a poměrně nepohodlně specifikovat, v čem může tato změna spočívat, kde se může vyskytovat. Nemůžeme zkrátka napsat pravidlo typu „shoduje-li se mnoho morfematických forem, na zbytku už nezáleží“. Takových drobných změn může nastat i víc naráz, a přesto člověk stále dobře vidí, že následující rámce jsou si hodně podobné:

  ACT(1) PAT(o+6)^{opt} ADDR(3)^{opt}
  EFF(4, aby+V, ať+V, cont, zda+V, že+V)
  MANN^{typ} **MEANS(7)**^{typ}

  ACT(1) PAT(o+6)^{opt} ADDR(3)
  EFF(4, **SPEECH,** aby+V, ať+V, cont, **jestli+V,** zda+V, že+V)
  REG(LN)^{typ}

  Valenční rámce jsou téměř stejné, ačkoli ADDR je jednou obligatorní a podruhé fakultativní, ačkoli morfematické formy jednoho EFF jsou podmnožinou druhých, ačkoli typické funktory jsou pokaždé jiné.

  Otázka zní, jakým jiným formalizmem lze tuto vágní podobnost elegantně popsat. Obáváme se, že určitá neobratnost zvoleného řešení regulárními výrazy – ačkoli částečně souvisí s předchozím bodem – by pravděpodobně byla vlastní i většině jiných metod.
- **Duplikace pravidel:** V případech, kdy nezáleží na vlastnostech jednotlivého slovníku, ale kdy potřebujeme říci například „rámce se liší o tento funktor“, je přesto nutné specifikovat, zda chybí ve VALLEXu, nebo v PDT-Vallexu. Pokud může chybět v jednom i ve druhém, je nutné uvést obě pravidla (ať už doslova, nebo třeba s nějakým příznakem k automatickém zdvojení).
- **Substituce:** Pokud víme, že např. funktor ADDR ve VALLEXu často odpovídá funktoru BEN v PDT-Vallexu, rádi bychom tolerovali rozdíl spočívající pouze v této záměně. To ale také regulární výrazy neumožňují. Je nutné například všechny výskyty nejprve nahradit a teprve poté rámce porovnávat.

Případně porovnávat rámec původní a posléze i substituovaný. Jde však již o zásah předcházející práci s regulárními výrazy.

Ze zpětného pohledu se může zdát, že volba samotných regulárních výrazů byla v určitém směru svazující. Přesto to považujeme za užitečné rozhodnutí. Regulární výrazy jsou standardní prostředek se známými vlastnostmi a obrovskou část práce udělaly za nás. Zmíněné omezení na počet prvků seznamu, které musíme jednotlivě vyjmenovat, nemá praktický dopad, neboť můžeme toto maximum dopředu nastavit dostatečně vysoké podle dat ve slovníku tak, aby žádný seznam nebyl delší. Většinu potíží jsme tedy vyřešili v rámci systému, který je zde popisovaný.

Kdybychom se rozhodli pro jiné řešení, nevyhnuli bychom se nutnosti zavést nějaký formalismus pro popis pravidel.[40] A každý formalismus vykazuje v některých oblastech ve srovnání s přirozenou řečí neobratnost či přílišnou „upovídanost“, která komplikuje zápis pravidla a zamlžuje jeho opětovné pochopení.

Z těchto důvodů jsme se přiklonili k regulárním výrazům a jejich rychlé a bohaté implementaci v Perlu.[41]

5.1.8 Opakované pouštění na novějších datech

Výhodou propojování slovníků, které pracuje automaticky, je možnost pouštět takovou proceduru opakovaně, což v našem případě bude nutné. Oba slovníky, VALLEX i PDT-Vallex, se nadále vyvíjejí, rozšiřují, doplňují i opravují, což probíhá v obou slovnících odděleně z důvodů odlišné koncepce a pravidel na tvorbu slovníkových hesel. Za čas bude tedy nutné zopakovat celou proceduru a provázat dvojici nových verzí slovníků.

Pokud do té doby provedeme nějaké ruční zásahy (opravy chybně spojených rámců, doplnění některých chybějících propojení), zaznamenáme tyto změny také odděleně. Výsledky budoucích propojení těchto slovníků budou vždy opraveny ještě o tyto změny. Protože se ve slovnících mohou měnit i stávající rámce (opra-

[40] Použít pravidlový přístup bylo prvotní rozhodnutí. Kdybychom zvolili statistické metody, které by se z trénovacích dat naučily, které lexikální jednotky se mají propojit, museli bychom jim dodat trénovací data a připravit rysy, které by tyto metody vyhodnocovaly.

[41] Spíše než nahradit celý systém by bylo v případě potřeby možné jej různě rozšiřovat a nabalovat na něj další funkcionality. Bylo by například možné tento systém vně modulu s regulárními výrazy rozšířit tak, aby umožnil zapisovat vedle pravidel také pravidlové šablony, z nichž by potom zvláštní generátor pravidel vytvořil desítky či stovky pravidel. Teprve tato pravidla by byla již standardně zpracována jako regulární výrazy.
Dále jsme mohli více využívat předzpracování. Podobně jako jsme funktory a formy seřadili, mohli jsme také některé funktory nahrazovat jinými apod.
Obojí by však samozřejmě zvyšovalo komplikovanost zápisu pravidel (přibývalo by syntaktických pravidel, co a jak lze zapsat) i kódu na jejich zpracování.

vovat, nebo i slučovat, či rozdělovat na více), ke každému záznamu o ručním zásahu si kromě identifikátoru lexikální jednotky uložíme také valenční rámec. Opravu potom provedeme pouze v případě, že odchylka mezi uloženým a novým rámcem nebude příliš velká.

5.1.9 Závěr

Popsali jsme postup, jakým propojujeme lexikální jednotky z VALLEXu s odpovídajícími valenčními rámci z PDT-Vallexu a naopak. Systém porovnává více kritérií, z nichž nejdůležitější je lexikální repertoár v glosách a příkladech a porovnání valenčních rámců ručně připravenými pravidly, která specifikují, které dvojice rámců lze považovat za dostatečně podobné. Na základě vzorku 200 sloves (což činí 6 % všech společných sloves) odhadujeme výsledné F_1 score pro celé slovníky na 77 %, což považujeme za uspokojivý výsledek. Byla vydána nová data VALLEXu, která pro úspěšně namapované lexikální jednotky obsahují informaci o protějšku z PDT-Vallexu; tato data byla rovněž zveřejněna v online verzi slovníku VALLEX, odkud vede od LU odkaz přímo do online slovníku PDT-Vallex.

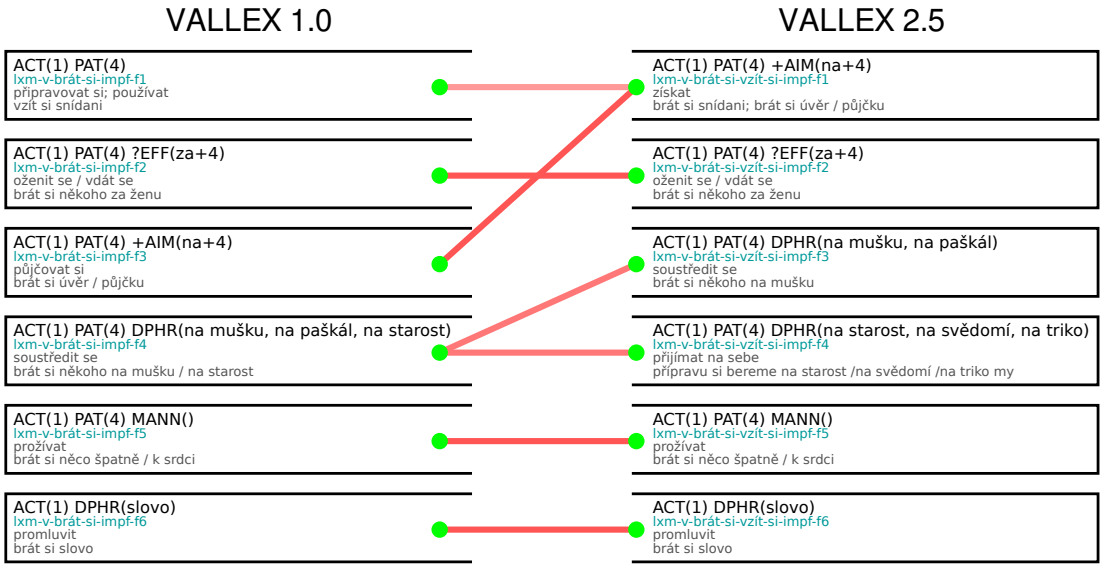
Tím jsme dosáhli též hlavního přínosu tohoto projektu, tedy napojení lexikálních jednotek VALLEXu na data PDT, o čemž pojednáváme v sekci 6.4.

5.2 Dvě verze VALLEXu

Dříve, než jsme propojili slovníky VALLEX a PDT-Vallex (popsáno v předchozí sekci 5.1), vyzkoušeli jsme si navržený postup a některá pravidla na jednodušším případě propojování dvou verzí jednoho slovníku: VALLEX 1.0 a VALLEX 2.5. (Výsledkem tohoto propojení bylo vydání verze 2.6, která se ničím jiným neliší.) V této práci jsme výklad obrátili a nejprve popsali výslednou komplexní úlohu. Nyní se tedy krátce vrátíme k jednodušší úloze, která komplexní předcházela, tedy k projektu Vallink nula.

Když jsme v sekci 4.2.1 mluvili o projektu VALEVAL, uvedli jsme, že při anotaci oněch zhruba 8 000 vět byly slovesům přiřazovány lexikální jednotky z VALLEXu 1.0. Ze slovníku, který oproti verzi 2.5 obsahoval méně slovesných lemmat, a pro ta lemmata, která obsahoval, pokrýval méně lexikálních jednotek. Při rozsáhlých úpravách slovníku při přechodu k verzi 2.5 (přidávání nových hesel, úpravy starých, slučování rámců, vytváření nových) nebyly zaznamenávány změny ani historie jednotlivých LU. Neexistovalo tedy mapování mezi starými a novými LU, čímž byla dočasně znehodnocena celá anotace v projektu VALEVAL. Ruční anotace je vždy nákladná záležitost, proto nebylo vhodné anotaci

opakovat, nýbrž raději automaticky nalézt ono mapování starých LU použitých k anotaci na nové LU z VALLEXu 2.5.



Obrázek 5.6: Správný výsledek automatického propojení odpovídajících si lexikálních jednotek mezi VALLEXem 1.0 a 2.5. Je zde vedle jednoduchých spojení totožných rámců vidět, jak se lexikální jednotky rozdělovaly na dvě i jak se slučovaly, to vše bez vztahu nového a původního identifikátoru LU.

Zpracování sloves ve staré a v nové verzi si bylo podobné, například pro sloveso *obracet* používané jako příklad v minulé sekci, ve verzi 2.5 jen přibyla sedmá LU a byl opraven valenční rámec třetí idiomatické LU s významem „zaměřovat“ z původního ACT(1) DPHR(zřetel,zájem,pozornost) DIR3 na nové ACT(1) PAT(k+3,na+4) DPHR(zřetel,zájem,pozornost). Nicméně i tak zde bylo nemálo rozdílů, jak ukazuje například obrázek 5.6.

Použili jsme tedy rozpracovaný základ automatické mapovací metody z předchozí sekce 5.1 a otestovali ji na jednodušším případě.

Formát. Už v tomto případě bylo potřeba převést VALLEX 2.5 do formátu SAMR, neboť VALLEX 1.0 ještě nebyl uspořádán na základě lexémů sdružujících vidové protějšky. Formát slovníku se změnil až mezi verzemi 1.0 a 2.5. Převod ze sdruženého „lexémového“ formátu B do formátu SAMR je rozhodně snazší nežli opačně.

Extrakce informací. Dvě verze VALLEXu jsou si pochopitelně podobnější i strukturou a typem informací. Z obou jsme tedy mohli například získat typická

valenční doplnění (která v PDT-Vallexu nejsou systematicky zachycovaná) nebo porovnávat lemmata nejen z příkladů, ale i z glos na obou stranách. Taktéž je zde značena idiomatická lexikální jednotka.

Pravidla pro porovnání valenčních rámců. Některé regulární výrazy použité jako pravidla pro porovnávání valenčních rámců se shodují s pravidly z minulé sekce (zejména ty, které jsou symetrické a aplikují se jednou na VALLEX a PDT-Vallex, podruhé na PDT-Vallex a VALLEX). Jiné jsme zde použít nemohli a další jsme naopak přidali. Například jsme počítali s tím, že docházelo k rozšiřování výčtu morfematických forem, takže několik pravidel očekávalo ve VALLEXu 2.5 stejné funktory s delším seznamem forem na vybraných pozicích.

Namapovali jsme takto všech 204 sloves z VALEVALu. Data byla následně použita pro propojení této nové verze VALLEXu 2.6 s textovými daty z projektu VALEVAL, což popisujeme v sekci 6.3.

5.3 VALLEX a FrameNet

Dalším projektem, o kterém je nutno se alespoň okrajově zmínit, ačkoli ještě stále probíhá a nemůžeme zde publikovat výsledky, je Vallink II, tedy propojení VALLEXu s FrameNetem (představení FrameNetu viz 4.6). Slovníky jsou si ve více ohledech mnohem vzdálenější, než byl VALLEX s PDT-Vallexem: FrameNet je spíše lexikální databáze, popisuje situace, které jsou evokovány lexikálními jednotkami a navzájem propojeny hustou sítí sémantických vztahů; je to *anglický* jazykový zdroj; pojetí rámcové sémantiky je více sémantické, a tudíž se liší od valenční teorie FGP založené na syntaxi.

FrameNet obsahuje mnohem více sémantické informace, která ve VALLEXu dosud chybí. V něm nejsou valenční rámce shlukovány do obecnějších situací platných pro větší množství LU, ani nejsou tyto nijak sémanticky provazované. Propojení těchto dvou lexikografických zdrojů bude tudíž pro VALLEX velmi přínosné.

Naším cílem je propojit nejprve lexikální jednotky a valenční rámce z VALLEXu se sémantickými rámci (semantic frames, SF) z FrameNetu, posléze v rámci takto vzniklých dvojic propojit jednotlivá valenční doplnění s elementy rámce (frame elements, FE). K tomu využijeme relaci Inheritance. Prvním krokem však musí být samotný překlad: vytipování anglických slovesných lemmat odpovída-

jících českým lexémům.[42] Teprve anglická slovesa „evokují“ sémantické rámce. Těch může být více, neboť každé lexikální jednotce z lexému může odpovídat jiný překlad či více překladů. Tato přeložená slovesa pak opět mohou evokovat více SF. Z nich musíme umět vybrat ty vhodné rámce.

↪ *Příklad:* Lexikální jednotce $odvést^V_1$ (s významem „vedením dopravovat *někam*“) odpovídají ve FrameNetu podle anotace, kterou přináší Benešová, Lopatková a Hrstková (2008), tři SF: Cotheme (přes překlady *conduct* i *escort*), Redirecting (přes *divert*) a Bringing (přes *take*). Druhé LU $odvést^V_2$ (s významem „vedením dopravovat *pryč*“) odpovídají hned čtyři rámce: přes překlad *take away* i přes *take out* jsou to Removing a Bringing, dále sémantický rámec Cotheme (přes *lead away*) a Cause_motion (přes *drag away*). Pokud ale sloveso *odvést* pouze „naslepo“ přeložíme všemi způsoby, potom seznam SF, do nichž možné překlady spadají, bude obsahovat minimálně 18 položek.[43] Například pro zmíněné $odvést^V_2$ v něm budou čtyři správné SF a třináct chybných.

Sémantické rámce z FrameNetu ⟷ lexikální jednotky z VALLEXu

Jak je tedy patrné, propojování různojazyčných zdrojů je náročné už v samotném počátku, kdy se počet možností překladem násobně zvýší. Pro nalezení těch správných SF z mnoha nabízených využijeme mj. sémantické třídy uvedené ve VALLEXu a hierarchický systém rámců ve FrameNetu uspořádaný relací Inheritance.

K tomu nám poslouží výstupy z projektu ruční anotace některých LU na SF FrameNetu (Benešová et al., 2008; Kettnerová, Lopatková a Bejček, 2012a). Ve zmíněných článcích je zpracována množina sémantických rámců odpovídajících jedné každé z osmi vybraných sémantických tříd VALLEXu; tuto množinu budeme značit SF(třída). Tyto SF jsou výsledkem mapování českých LU z jednotlivých sémantických tříd do SF, které byly následně zobecněny právě pomocí hierarchie Inheritance. Všechny zobecněné SF se nalézají v horních „patrech“ této hierarchie a zastupují mnohé nižší SF, které byly také výsledkem mapování. Po tomto

[42] Chceme se striktně držet přístupu přes anglická slovesa a jejich příslušnost do SF. Nepovolujeme přiřadit lexikální jednotce sebevhodnější sémantický rámec, pokud do něj ve FrameNetu není přiřazen některý z možných překladů českého slovesa.

[43] Vycházíme z ruční anotace části lexikálních jednotek a jejich analýzy, viz sekce 4 citovaného článku. Čtyři ze sedmi lexikálních jednotek slovesa *odvést* byly přeloženy osmnácti anglickými ekvivalenty. Z nich sedm se ve FrameNet vůbec nevyskytuje, zbylé evokují zmíněných 18 SF.

zobecnění autoři získali 81 SF pro osm sémantických tříd (communication, mental action, social interaction, psych verbs, exchange, motion, transport a location).[44]

Postup, který používáme, je znázorněn na obrázku 5.7. Zvolené české lemma se vyskytuje v několika LU, každá má svůj valenční rámec. Zaměříme se na mapování lexikální jednotky č. 2. Lemma přeložíme automaticky do angličtiny[45] ve významu odpovídajícím LU_2 a získáme celou řadu anglických lemmat, která evokují stejný, nižší nebo vyšší počet SF. Vybereme jen ty SF, které sémanticky odpovídají zvolené české LU_2, tedy ty, které jsou obsaženy v množině SF(sémantická třída$_2$), nebo jsou od nich odvozené relací Inheritance (na obrázku znázorněno šedým kruhem). Zvolené CZ LU_2 tedy odpovídají EN $LU_{1.2}$, EN $LU_{2.2}$ a EN $LU_{3.1}$. (Opakujeme pro případné další LU náležející do jiných sémantických tříd.)

Z nich dále vybereme na základě počtu (a částečně i typu) elementů rámce. Tzv „core“ elementy představují ty členy, které jsou koncepčně nezbytné pro popisovanou situaci (pro daný SF) a jejichž kombinace je pro SF charakteristická (Ruppenhofer et al., 2006). Na základě pozorování víme, že je častá korespondence mezi tzv. „core“ elementy sémantického rámce FrameNetu a tzv. „valenčním rámcem v užším smyslu slova“ VALLEXu (tedy aktanty plus obligatorní volná doplnění). Porovnání počtu „core“ elementů a členů valenčního rámce v užším smyslu slova nám umožní vyloučit další falešné kandidáty. Částečně můžeme také využít informaci o tom, které valenční doplnění z VALLEXu často odpovídají kterým elementům z FrameNetu. Základem systému tedy opět bude framework navržený pro porovnávání valenčních slovníků v sekci 5.1 VALLEX a PDT-Vallex.

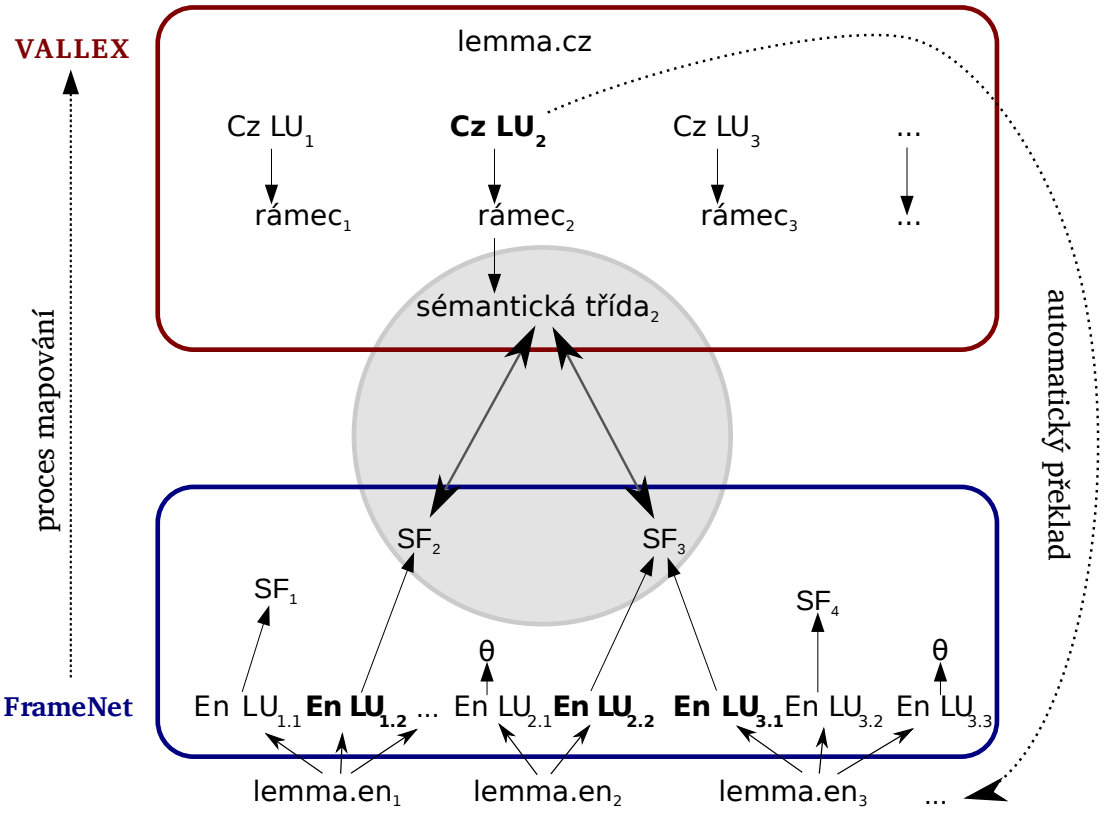
Tak bude na každou LU z VALLEXu (která má přiřazenou sémantickou třídu) namapována sada odpovídajících sémantických rámců z FrameNetu, které jsou evokovány anglickými ekvivalenty lemmat z české LU.

Elementy rámce z FrameNetu ⟷ valenční doplnění z VALLEXu

V druhé části experimentu valenčním doplněním (opět pouze z valenčního rámce v užším smyslu slova) namapovaných lexikálních jednotek přiřazujeme „core“ elementy rámce, neboť tyto skupiny si prototypicky odpovídají. Provedli jsme ruční analýzu a sestavili seznamy FE, které odpovídají jednotlivým valenčním

[44] Například pro třídu transport, ze které je také sloveso *odvést*, obsahuje seznam těchto šest sémantických rámců:
SF(transport) = {Cause_motion, Bringing, Cotheme, Filling, Firing, Releasing}.

[45] Pomocí slovníku `http://slovnik.zcu.cz`.



Obrázek 5.7: Schéma propojování lexikální jednotky z VALLEXu se sémantickými rámci z FrameNetu. Dole jsou překlady českého lemmatu odpovídající anglickým LU. Více lexikálním jednotkám může odpovídat stejný sémantický rámec (zde $LU_{2.2}$ a $LU_{3.1}$), některé ve FrameNet nemusí být zahrnuty vůbec (např. $LU_{3.3}$).

doplněním z VALLEXu.[46] Na základě této korespondence se pokoušíme nalézt správné dvojice valenční doplnění – element rámce.

Tento projekt zde uvádíme pro doplnění spektra možností, jaké lexikální databáze a slovníky je možno provazovat a jakou škálu informací k tomu používat. Jak jsme uvedli, výsledky jsou zatím pouze předběžné a projekt stále pokračuje.

[46] Například pro funktor EFF je to jen několik možných FE, například ve třídě mental action je to vždy CONTENT, ve třídě communication jedna z možností MESSAGE, CATEGORY, nebo PHENOMENON, či ve třídě location buď WHOLE, nebo CONFIGURATION.

5.4 PDT-Vallex a FrameNet

Na závěr této kapitoly si dovolíme zmínku o synergickém efektu, kdy budeme moci po propojení VALLEXu a FrameNetu (sekce 5.3) propojit bezpracně také PDT-Vallex s FrameNetem, neboť VALLEX a PDT-Vallex už propojený máme (sekce 5.1). Díky tomu získáme částečnou anotaci dat PDT sémantickými rámci z FrameNetu.

6

Propojování slovníku s textem

Vypovídací hodnota slovníku se značně zvýší, pokud se podaří jevy, které zachycuje, nalézt a vyznačit v datech (v textu), a tím je exemplifikovat na reálných textech. Takovéto propojení slovníku a dat přináší nové prostředky pro práci lexikografů i lingvistů. Pro statistické metody zpracování přirozeného jazyka jsou pak takovéto jazykové zdroje přímo nezbytné.

↪ *Příklad:* Mějme ve slovníku u slova *klín* jen tyto dva významy: 1. zašpičatělý plochý kus dřeva, kovu aj. sloužící k rozštípnutí ap.; 2. přední část těla od pasu ke kolenům u sedící osoby (SSČ). Chceme-li je rozlišovat v textu automaticky, jsou dva základní přístupy:

- Můžeme se pokusit navrhnout pravidla, která rozdíly vystihnou. Snadný je třeba genitiv, ve kterém se slova odlišují: *spadlo mu jmění do klín**a*** vs. *marně bije palicí do klín**u***. Dále (už s menší spolehlivostí) můžeme použít nejbližší okolí slova, kdy *posadit na klín* bude téměř vždy představovat význam č. 2. apod. Na přípravu a ověřování takových pravidel však potřebujeme anotovaný korpus (leda bychom chtěli všechny příklady procházet manuálně a rozhodovat o jejich významu – tedy provádět ad hoc anotaci).
- Druhý přístup je statistický a vychází právě z ručně anotovaného vzorku dat – „trénovací data“. Ta jsou v takovém případě nenahraditelná, neboť z nich se systém „naučí“ klasifikovat, v kterém kontextu a v jakém tvaru se vyskytují oba významy.

↪ *Příklad:* Pokud jsou fráze ve slovníku (např. hovorové *klíďo*, *klíďo brďo* a *klíďo píďo*) uvedeny bez dalšího propojení s daty, jen těžko zjistíme, že jedna z nich se téměř nepoužívá (o přesných frekvencích nemluvě), a už vůbec nemůžeme sledovat, v jakých kontextech je která z nich použitelná a zda se mírně neliší jejich význam.

Navíc proces propojování slovníku s (reprezentativními) texty (tedy anotace) může velmi pomoci slovníku samotnému. Umožňuje například zjistit, že slovníková hesla mají nízké pokrytí, nebo se při něm dokonce mohou objevit chybějící položky s častým výskytem v datech; potom je možné slovník vhodně dodatečně rozšířit. Aplikace – do té doby třeba i vágních – kritérií odlišujících jednotlivé po-

ložky slovníku prospěje konzistenci, přesnosti a použitelnosti slovníku průběžnou úpravou kritérií nebo i samotných hesel.

Na anotaci řady známých, používaných korpusů se dá pohlížet jako na propojení jazykového jevu ve slovníku s textem. Jmenujme pro češtinu ty, se kterými budeme pracovat (za každým jazykovým zdrojem uvádíme v závorce sekci této práce, která se jím zabývá):

1. korpus PDT ve verzi 2.5 (sekce 3.4) obsahující ruční anotaci víceslovných výrazů ze slovníku SemLex (4.9) v projektu Lexemann (sekce 6.1.3 Ruční anotace víceslovných výrazů),
2. korpus PCEDT (3.5) a ruční anotace entit od společnosti BBN (6.2 PCEDT a entity BBN), ačkoli v tomto případě by „slovník“ obsahoval jen desítky kategorií,
3. část korpusu SYN 2000 (3.1) a ruční anotace valence vybraných sloves ze slovníku VALLEX 1.0 v projektu VALEVAL (4.2.1),
4. korpusy PDT (3.3) a PCEDT (3.5) a ruční anotace valence pomocí slovníku PDT-Vallex (4.3) jako integrální součást projektů PDT 2.0 a PCEDT 2.0.

S využitím těchto jazykových zdrojů mohou vznikat nová (anotovaná) data do značné míry bez dalších ručních zásahů. V následujících sekcích popíšeme vznik těchto automatických anotací korpusů:

6.1 anotace víceslovných výrazů ze SemLexu — v korpusu ČNK
(automaticky na základě dat z bodu 1)

6.2 kontrola konzistence anotace víceslovných entit BBN — v korpusu PCEDT
(automaticky na základě bodu 2)

6.3 anotace valence vybraných sloves z VALLEXu — v části korpusu ČNK
(vychází z bodu 3 a z propojení slovníku VALLEX ve verzích 1.0 a 2.5 popsaného v sekci 5.2)

6.4 anotace valence sloves z VALLEXu — v korpusech PDT a PCEDT
(vychází z bodu 4 a z propojení slovníků VALLEX a PDT-Vallex popsaného v sekci 5.1)

6.1 ČNK a SemLex

V této sekci představíme projekt identifikující automaticky víceslovné výrazy (VV) ze slovníku v textu. K tomu použijeme slovník SemLex představený již v sekci 4.9), a texty z Českého národního korpusu. Výsledky projektu byly publikovány v Bejček, Straňák a Pecina (2013). Autor této práce měl na starosti většinu technické části projektu a podílel se na návrhu metod.

6.1.1 Úvod

Víceslovné výrazy (VV), jež (ze své definice) vždy nesou nějaký idiosynkratický rys, se chovají odlišně než pravidelně tvořené syntaktické konstrukce. Samozřejmě se ani VV nechovají jako homogenní skupina, jejich vlastnosti se jeden výraz od druhého značně odlišují. Pro naprostou většinu úloh zpracování přirozeného jazyka je vytipování VV nezbytně nutným krokem.

Bez něj by se například *„stahuje kalhoty, když brod je ještě daleko“* překládalo do všech jazyků doslovně, po jednotlivých tokenech; bylo by těžké ve vyhledávači nahradit výrazy *„ministerský předseda“*, *„předseda vlády“* či *„první ministr“* synonymem *„premiér“* (případně dokonce rozdělený výraz *předseda české vlády* nahradit výrazem *český premiér*); při analýze textu pro extrakci informací (*information extraction*) by frázové sloveso *„look forward to“* zvýšilo riziko, že věta bude chybně analyzována jako hovořící o *hledění někam*, neboť se v ní vyskytuje sloveso *look*.

Jedná se pochopitelně jen o motivační příklady, oblast NLP je různorodá a široká, její úlohy lze řešit mnoha různými postupy, které vždy explicitní vyznačení VV obsahovat nemusí. Nelze tedy jednoduše tvrdit, že by výstupy bez identifikace VV byly vždy špatné. Například některé statistické metody se o identifikaci postarají nativně aniž by řešily, co je VV a co není. Ovšem pro jiné metody je vyhledání bloků, které se mají zpracovávat naráz, zásadní – a nakonec i statistickým metodám může pomoci předzpracování, které např. „opraví“ tokenizaci tak, že sloučí VV do jediného tokenu. Eryigit, Ilbay a Can (2011) provedli řadu experimentů s víceslovnými výrazy zpracovanými před fází závislostní syntaktické analýzy. Zjišťují, že plošné sloučení víceslovných výrazů do jednoho tokenu sice analýze dokonce škodí, neboť „ředí“ data zaváděním nových neznámých tokenů. Ovšem zaměření na vybrané typy výrazů, které parser neodkáže sám zpracovat dostatečně dobře, naopak celkovou úspěšnost zvyšuje. Řada dalších autorů reportuje zlepšení překladového systému zapojením identifikace VV, jmenujme např. Ren et al. (2009); Pal, Chakraborty a Bandyopadhyay (2011); Bouamor, Semmar a Zweigenbaum (2012).

Naším cílem není kompletní vyhledání všech (potenciálních) VV, natož všech signifikantních kolokací. Omezujeme se na vyhledání těch VV, které máme uvedeny ve slovníku (mluvíme proto o *identifikaci* VV **v textu**, která stojí v protikladu k úkolu označovanému anglickým termínem *acquisition*). Propojování *slovníku* s daty je ostatně tématem celé této kapitoly. Propojení výskytů VV se slovníkem (což jiné přístupy k vytipování VV nevyužívají – často ani nemají dopředu stanoveno, které výrazy budou klasifikovány) přináší ovšem řadu výhod, neboť ve slovníku mohou být uloženy další údaje o výrazu: jeho synonyma, tzv. *ukotvení*

(*grounding*)[1] pojmenovaných entit, vzájemné zanoření a vůbec odkazy na jiná hesla slovníku, slovní druhy atp.

Také neřešíme problém doslovného vs. frazeologického užití VV. Domníváme se totiž (výsledky projektu a naše zkušenosti to potvrzují, ale pečlivější výzkum nemáme k dispozici), že doslovné užití je, přinejmenším v češtině, poměrně řídké (neboť je v zájmu autora i příjemce textu, aby nevznikalo nedorozumění) a dochází-li k němu, jedná se zpravidla o jazykové hříčky, záměrné dvojsmysly, či jazykový humor. Důvodem může být velká volnost češtiny (oproti angličtině např. v slovotvorbě, slovosledu či morfologii), která umožňuje v případě potřeby doslovného významu daný obrat pozměnit, čímž přijde o svůj frazeologický význam.[2]

Představovaný projekt Lexemann, jak bylo řečeno, identifikuje jen ty VV, které jsou uvedeny ve vstupním slovníku. Slovníkem se řídí výhradně: metoda v této základní podobě není stochastická, nehledí na okolí ve větě ani na relativní frekvenci. Proto nejprve přestavíme manuální anotace VV v PDT 2.5, které probíhaly nad tektogramatickou rovinou a při nichž vznikl zmíněný slovník (v sekci 6.1.2 definujeme pojmy a v sekci 6.1.3 popíšeme anotovaná data), a také automatický tektogramatický parsing využívající systém Treex (sekce 6.1.4). Poté přejdeme k teoretickým východiskům metody (sekce 6.1.5) a k samotnému automatickému rozpoznávání VV, které hledá známé vzory tektogramatických podstromů v automaticky parsovaných datech, a ukážeme, jaké úspěšnosti dosahuje (sekce 6.1.6). Výsledky rozebereme a zdůvodníme 6.1.7.

6.1.2 Víceslovné výrazy, lexie, pojmenované entity

Za VV považujeme v popisovaném projektu jednak víceslovné *lexie*, jednak víceslovné *pojmenované entity*. Nyní se pokusíme vymezit, jak tyto pojmy chápeme. Jsme si plně vědomi složitosti problematiky, neustálenosti některých pojmů a obtížné definovatelnosti jiných, což vedlo k celé řadě různých definic. Neklademe si tedy za cíl (a pro účely této práce není ani nutné) tuto oblast rozpracovat po teoretické stránce. Potřebujeme pouze čtenáři předat představu toho, s čím projekt, na nějž navážeme, pracoval.

[1] Groundingem, neboli ukotvením rozumíme svázání výrazu s jeho denotátem. *Náměstí Jiřího z Poděbrad* tak může mít ve slovníku uvedeny své GPS souřadnice a zanořený *Jiří z Poděbrad* například odkaz na patřičný článek ve Wikipedii. Viz též sekce 7.2.2.

[2] Např. *„Vilém už dávno **od**hodil pušku do žita.“*, *„Průměrná žena **zastane** v domácnosti většinu prací.“* (přerušení VV *žena v domácnosti* a tím zrušení jeho významu), *„chvála boh**ovi**“*. Také je mnoho VV, které se těžko dají použít v jiném významu, protože vlastně nic jako *původní* význam nemají, např. *„brigáda rychlého nasazení“*, nebo *„rostlinný tuk“*.

Lexie

Nejprve se podíváme na definici pojmu lexie samotné (bez vztahu k VV). Lexii chápeme v souladu s českou tradicí stejně jako Filipec a Čermák; tento pojem odpovídá Cruseho lexikální jednotce, případně Kempsonovu lexému.

Filipec a Čermák (1985) i sám Filipec (1994) chápou *lexii* jako slovo nebo slovní spojení v jednom konkrétním významu. Tak ho budeme chápat v celé práci.

Filipec a Čermák, str. 21, definují lexii jako termín ekvivalentní k *monosémickému lexému*. Také je, jak uvidíme, synonymní s jejich *základní lexikální jednotkou*. Doslova pak jejich definice, vycházející z lexému, vypadá takto: Termín lexém je „synonymní s termínem lexikální jednotka, ale je ještě dále diferencován. […] Lexém může mít jeden význam (monosémický lexém) nebo dva i více významů (polysémický lexém). Lexikální jednotka jako polysémický lexém je čtyřstupňový útvar zahrnující tolik různých monosémických základních lexikálních jednotek, lexií (v schématu L_1-S_1, L_1-S_2), kolik má různých významů. […] Polysémický lexém v jeho zahrnující a nadřazené funkci označujeme jako *hyperlexém*. Hyperlexém je jednotka asymetrická: jedna reprezentativní lexikální forma může mít dva i více významů; lexie je jednotka symetrická. Nepominutelnou jednotkou je i *alolex*, tj. exemplář lexie v textu a ve výpovědi.“ (Filipec a Čermák, 1985, str. 28, 29)

Filipcova lexie zhruba odpovídá pojmu *„lexical unit“* u Cruseho (1986). Cruse definuje *lexikální jednotku* jako spojení lexikální formy (což je soubor různých potenciálních tvarů jednoho slova) a jediného významu. Je zde však patrně rozdíl ve značné jemnosti, s jakou Cruse vyčleňuje další a další lexikální jednotky a pracuje s jinak definovaným pojmem *lexém*$_2$: „Senses [i.e. the meaning aspects of lexical units] need to represent unitary ‘quanta’ of meaning, but they do not need to be finite in number. […] A lexeme, on the other hand, may well be associated with indefinitely many senses, but the set of lexemes must be finitely enumerable.“

Výrazu „bunda“ ve větě „Tahle bunda se mi líbí.“ pak přiřazuje odlišné lexikální jednotky pro významy „tato konkrétní bunda“, „bunda tohoto typu“, „bunda této barvy“, „bunda tohoto střihu“ atd.

Cruse dále odkazuje ke Kempsonovu pojmu lexém (Kempson, 1977), který je velice blízký Cruseho lexikální jednotce a tudíž zřejmě též Filipcově (a tedy i naší) lexii.

V této práci využíváme již hotový slovník, proto dál tuto problematiku nebudeme rozebírat. Zavedli jsme pouze pojmy, z nichž budeme nadále používat lexii.

Víceslovné lexie

Dosud jsme se věnovali pojmu lexie. *Víceslovná* lexie se pochopitelně musí skládat z více než jednoho slova,[3] ale jak rozhodnout, zda nějaké sousloví *je* lexie, tedy zda tvoří jednu významovou jednotku? Podle našeho názoru nelze stanovit ostrou hranici.[4] Sestavili jsme tedy několik kritérií, jejichž naplnění indikovalo, že se o víceslovnou lexii jedná a vyškolený anotátor by ji měl v textu označit.[5] Tato kritéria byla formulována pomocí následujících vlastností jednoslovných lexií (případně řady jednoslovných lexií). Čím více z těchto vlastností výraz porušoval, tím jistější si anotátor byl, že se jedná o lexii víceslovnou:

kompozicionalita Toto bylo nejdůležitější kritérium pro určení víceslovných lexií a v řadě případů vystačilo samo o sobě. *Kompozicionální* jsou taková spojení, jejichž význam lze dovodit z jejich částí. Nekompozicionální spojení je přesvědčivý kandidát na víceslovnou lexii.

↪ *Příklad:* Můžeme si tedy představit pokusnou osobu, která zná všechna slova jazyka, zná i gramatická pravidla a je postavena před nový výraz *kuchyňský nůž*. Samozřejmě chápe, že to je nůž související s kuchyní. Je ale vůbec možné z jazyka zjistit, zda se jedná o ostrý nůž na krájení surovin na prkénku, nebo o příborový nůž na krájení hotového jídla, jehož místo je taktéž v kuchyni? Dotyčná osoba to z jazyka neodvodí (ne s jistotou), proto tento výraz považujeme za víceslovnou lexii.

[3] Pro naše účely budeme dále pracovat s mírně zúženým pojetím víceslovné lexie. Protože jsme se orientovali na t-rovinu PDT, která reprezentuje pouze významová slova, bude se naše víceslovná lexie vždy skládat z více než jednoho **autosémantického** slova. Tímto rozhodnutím jsme přišli o část víceslovných lexií, například o některé sekundární předložky (*na rozdíl od* či *v důsledku*), ale v celkovém množství byl úbytek zanedbatelný.

[4] Následující řádky popisují poznatky z projektu Lexemann pod vedením Pavla Straňáka. Pokud je nám ovšem známo, tento rozhodovací proces nebyl dosud publikován.

[5] S anotátory jsme měli často schůzky, na nichž jsme rozebírali sporné případy, hledali univerzální řešení, zpřesňovali kritéria a zaznamenávali ukázkové příklady a jejich řešení. Pro určitou představu o víceslovných lexiích i o šíři problematiky uvádíme několik úvah, které probíhaly:
Víme, že *hlavní tah* má význam dopravní tepny. Jde to ale poznat z jazyka samotného? Nemohl by to být zásadní tah v šachové partii?
Integrovaná doprava je srozumitelná. A jak víme, *s čím* je integrovaná? Tato informace je už zakódovaná v samotném sousloví.
Má slovo *ekonomický* ve spojení *ekonomický náměstek* stejný význam jako v ostatních (běžných, neidiosynkratických) spojeních? Je to tedy „náměstek pro ekonomii“, nebo „na něm ředitel ušetřil“, stejně jako ušetříme na *ekonomickém čisticím prostředku*?
Můžu tvrdit, že ze spojení *nákladní vůz* nejde poznat, že se nejedná o žebřiňák (žebřiňák je jistě také vůz), nebo je ten význam skryt už v dílčích slovech, tedy v tom, že se jedná o $vůz_2$ (automobil) a nikoli o $vůz_1$ (vozidlo o čtyřech kolech tažené potahem, povoz) (SSČ)?

Ani u tohoto kritéria ovšem není situace jednoznačná. Například se můžeme ptát, kolik pragmatické informace může pokusná osoba do usuzování vnášet, zda zná ostatní víceslovné lexie, kde se vyskytují stejná lemmata, apod.

přeložitelnost Pokud by doslovný překlad (tedy po jednotlivých částech výrazu) do jiných jazyků nebyl srozumitelný, je to důvod domnívat se, že se jedná o víceslovnou lexii. (To může případně platit i o výrazech, u kterých se zdá, že je jejich význam plně poskládán z částí.)

↪ *Příklad:* Doslovný český překlad anglického výrazu *traffic light*[6] v češtině neexistuje a zřejmě by výraz **dopravní světlo* byl i obtížně srozumitelný. (Mohla by to být obrysová/potkávací světla automobilů? Může to být veřejné osvětlení dálnic?)

substituovatelnost části Za víceslovnou lexii považujeme i natolik fixní spojení, že není možné nahradit jeho část ekvivalentním slovem při zachování významu.

↪ *Příklad:* I pokud by se výraz *řádné studium* zdál srozumitelný, jeho ustrnulost si uvědomíme, pokud zkusíme varianty **správné studium*, **spořádané studium*, nebo **řádné učení*.

Podobně *státní činitel* → **státní konatel*.

přerušitelnost Posledním indikátorem, který zmíníme, je variovatelnost výrazu vložením nějakého rozvití. To je opět vlastnost, kterou některé ustrnulejší víceslovné lexie nepřipouštějí.

↪ *Příklad:* **dopravní*$_{(,)}$ ***závažný*** *přestupek*, **ministerstvo* ***českého*** *hospodářství*, ?*hodit přes* ***nemytou*** *palubu*.

Anotátoři používali i některé další – často i ad hoc – testy, které dokládají idiosynkracii výrazu. Konečné rozhodnutí museli učinit oni. Proto svou definici víceslovné lexie můžeme nejpřesněji formulovat jako *jakýkoli víceslovný výraz, který anotátor označil jako lexii*.

Pojmenované entity

Zavádění pojmů uzavřeme přiblížením pojmu *pojmenovaná entita*, (NE, *named entity*). Zmiňujeme je zde jen okrajově: projekt Lexemann s nimi pracoval v plné míře (víceslovných NE bylo v textu nalezeno více nežli víceslovných lexií), ovšem zde budeme pracovat pouze s víceslovnými lexiemi a pojem NE potřebujeme zejména k jejich vymezení. Kritériem pro anotaci víceslovných pojmenovaných entit bylo (vcelku přirozeně), aby výraz byl pojmenovaná entita a skládal se z více než jednoho slova.[7] Ačkoli snad intuitivně chápeme, co je pojmenovaná

[6] Česky správně *světelné signalizační zařízení*, lidově *semafor*.

[7] Opět, důležitý byl počet **autosémantických** slov. To nám sice podobně jako u lexií částečně bránilo v anotaci, zde např. restaurace *Na Šachtě*, politické strany *ODS* nebo ulice

entita, definovat ji je velmi náročný úkol, zcela mimo rozsah této práce. Navíc, jak upozorňuje Straňák (2010), snahy o jeho splnění většinou definují NE kruhem.

Pro účely této práce postačí jednoduše výčtem říci, že jsme jako pojmenované entity chápali:

- ustálené názvy, zejména názvy
  - osob či zvířat,
  - institucí,
  - objektů (např. knih, sloučenin, festivalů, botanických názvů),
  - geografické;
- strukturované údaje
  - bibliografické,
  - adresy;
- časové údaje
- a také částečně mimojazykové
  - číselné rozsahy,
  - výrazy z jiného jazyka.

6.1.3 Ruční anotace víceslovných výrazů

V této sekci stručně shrneme důležité informace o projektu Lexemann (pro zevrubnější popis viz zejména disertační práci Straňák, 2010 a článek Bejček a Straňák, 2010, příp. Bejček, Straňák a Hajič, 2009). V rámci tohoto projektu byla vytvořena úplná anotace víceslovných výrazů na celé tektogramatické rovině PDT 2.0.[8]

V průběhu tří let celkem pět anotátorů zpracovalo všechny texty PDT 2.0, které byly anotované až na tektogramatickou rovinu. V textech označili všechny víceslovné výrazy (lexie i NE). Celkem zvažovali zhruba 675 000 t-uzlů[9] (z toho

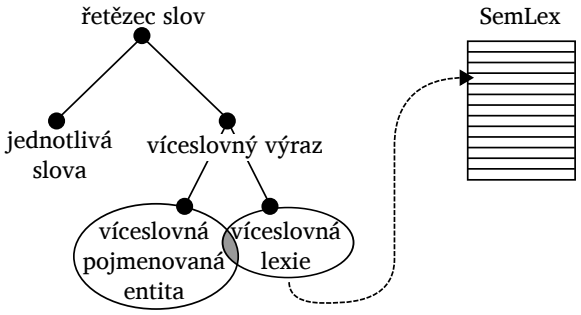
Ve Smečkách, ale omezení nebylo citelné. Samozřejmě už bychom se zabývali restaurací *Na Nové Šachtě*, což je její dnešní název, plným názvem *Občanské demokratické strany* nebo *Restaurací Ve Smečkách*, která má ve svém názvu autosémantická slova dvě.

[8] Tato data jsou k dispozici ke stažení na adrese `http://ufal.mff.cuni.cz/lexemann/mwe/` nebo jako součást PDT 2.5 (Bejček et al., 2012 a `http://ufal.mff.cuni.cz/pdt2.5`); obojí pod svobodnou licencí Creative Commons BY-NC-SA 3.0.

[9] Zvažovali, zda je uzel součástí nějakého VV; které další uzly do VV patří (tedy jaký je jeho rozsah) a o jaký typ pojmenované entity se jedná, nebo jaké slovníkové heslo odpovídá této lexii. (To může být obtížné rozhodnout třeba u elipsy: existuje *„čistička vod“*, nebo se vždy jedná o *„čističku odpadních vod“*? Mají být ve slovníku dvě hesla, nebo stačí to druhé a kratší případ je jen elipsa?) Toto víceúrovňové rozhodování bylo potřeba vzít v úvahu při počítání mezianotátorské shody (Bejček, Straňák a Schlesinger, 2008) a ještě se k němu vrátíme na konci sekce 6.1.6 na straně 163 při vyhodnocování úspěšnosti automatické procedury.

51 % dat zpracovali paralelně dva anotátoři a dalších 12 % tři anotátoři), které odpovídají 713 000 slov či 833 000 tokenů.

Jak jsme řekli v minulé sekci, za VV považujeme veškeré *pojmenované entity* a všechny *lexie* (frazémy, analytické predikáty, idiomy, termíny apod.) – pokud jsou (jedny nebo druhé) složené z více než jednoho autosémantického slova, viz obrázek 6.1. Pro snadnější čitelnost budeme nadále výraz *víceslovná lexie* používat zjednodušeně pouze pro takové VV, které nejsou pojmenovanou entitou. Vznikne tedy opozice mezi víceslovnou pojmenovanou entitou a víceslovnou lexií. Právě víceslovné lexie nás budou zajímat, neboť anotátoři je (kromě vyznačení v textu) také vkládali do slovníku SemLex (viz sekci 4.9) vytvářeného právě za tímto účelem.



Obrázek 6.1: Schéma znázorňuje vztah víceslovných lexií k ostatním pojmům. Přechod od víceslovné pojmenované entity k víceslovné lexii je pozvolný a mnoho hraničních případů je obtížně rozhodnutelných (například názvy zákonů).

V této sekci budeme často místo *víceslovná lexie* psát pouze *lexie* ve stejném významu (zejména v kontrastu k pojmenované entitě). Podobně budeme v kontextu slovníku SemLex zaměňovat *(víceslovná) lexie* a *víceslovný výraz* (VV) – z prostého důvodu, že v SemLexu jsou uloženy pouze ty VV, které jsou víceslovnými lexiemi, a proto mluvě o slovníku oba pojmy splývají.

Takto byla anotována veškerá data t-roviny PDT. Na více než polovině dat byla anotace paralelní. Z ní jsme získali „gold data“ poloautomatickým způsobem řídícím se typem anotátorské neshody – tedy zda se výraz jednoho a druhého anotátora nějak překrývá. Pravidla byla následovná:

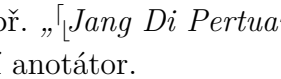
- Pokud se anotátoři shodli, víceslovný výraz byl zachován. V lepším případě se anotátoři shodli i na označkování VV a mohla být zachována značka. Pokud se neshodli, pro PDT 2.5 a PDT 3.0 to nebyl velký problém, neboť jejich součástí není ani SemLex ani odkazy do něj: víceslovné lexie jsou pouze opatřeny značkou LEXEME (míněno monosémický lexém). Odkazy na dvě shodné položky slovníku stejně jako na dvě rozdílné byly nahrazeny touto značkou LEXEME.[10] Vzhledem k vágní hranici mezi pojmenovanou entitou a lexií dopadly i dvě různé NE stejně: nejsou-li anotátoři zajedno, o jakou NE se jedná, uložíme v budoucnu celý výraz do slovníku. Jedině při rozporu NE versus lexie volíme specifičtější informaci, tedy typ NE.

- Pokud výraz označil pouze jeden anotátor, víceslovný výraz byl zachován. Pro neparalelní anotace je toto rozhodnutí samozřejmé. V paralelních anotacích se ukázalo, že je mnohem častější případ, kdy anotátor přehlédne VV, než kdy anotuje chybně nevýznamný řetězec slov, proto jsme tyto anotace zachovali do gold dat.

- Pokud jeden anotátor anotoval víceslovný výraz, který tvoří vlastní podmnožinu uzlů výrazu anotovaného druhým anotátorem (např. *„ústavní soud“* a *„předseda ústavního soudu“*), ponechali jsme výraz rozsáhlejší (*„předseda ústavního soudu“*). Tím se neztratí informace o přidaném slově (*„předseda“*) a ve slovníku lze v budoucnu specifikovat, že *„předseda ústavního soudu“* obsahuje slovníkové heslo (a tudíž VV) *„ústavní soud“*.

- Naopak jestliže jeden anotátor anotoval několik podmnožin výrazu druhého anotátora tak, že svou anotací plně pokryl daný rozsáhlejší výraz (např. *„⌈⌊hraniční přechod⌋ ⌊Lanžhot – Kúty⌋⌉“* nebo *„⌈⌊Staré purkrabství⌋ ⌊na Pražském hradě⌋⌉“*), zachovali jsme jednotlivé specifičtější anotace.

- V podobném případě, kdy ovšem podmnožiny nepokryly celý výraz (tedy se anotátoři navíc neshodli, které t-uzly jsou součástí víceslovných výrazů, např. *„⌈⌊Jang Di Pertuan Agong⌋ Sultan ⌊Azlan Shah⌋⌉“*), rozhodoval ručně třetí anotátor.



[10] Obě rozdílné anotace jsou si obvykle podobné, například slovníková hesla *smluvní lékař* a *smluvní lékař pojišťovny* pro dvouslovný výraz *smluvní lékař*, který jeden z anotátorů chápe jako elipsu. V „gold datech“ jsme se rozhodnout museli, přijali jsme tedy jako prozatímní řešení první z anotací. Zatímco ani jedna z anotací se nejeví jako očividná chyba, výběr libovolné z nich nám pochopitelně ublíží při evaluaci – zhorší výsledky naší automatické identifikace.

- V případě, že anotace dvou anotátorů se lišily a měly neprázdný jak průnik, tak oba doplňky (např. „⌊*novela* ⌈*zákona*⌋ *o* ⌊*dani z příjmů*⌉“), rozhodoval třetí anotátor.

Kromě „gold dat“ chceme mít stále možnost pracovat také s paralelními daty od jednotlivých anotátorů, proto jsme v SemLexu ponechali všechny víceslovné lexie, které byly při anotaci použity (bez ohledu na to, zda se dostaly právě popsaným způsobem do „gold dat“).[11] Tím jsme získali celkem 8 797 *typů*[12] víceslovných lexií (tedy položek v SemLexu), z nichž 8 502 zůstalo v „gold datech“ a bylo použito v 21 946 případech (*instancích*). Krom toho jsme získali také 26 448 instancí anotovaných víceslovných pojmenovaných entit, se kterými zde ovšem pracovat nebudeme. Viz též tabulka 6.1.

	typy (položky ve slovníku)	instance (výskyty v datech)	průměrné použití
víceslovné lexie	8797	21 946	2,5 ×
víceslovné NE	N/A	26 448	N/A
	(rozlišujeme 9 druhů)		(2 939 ×)

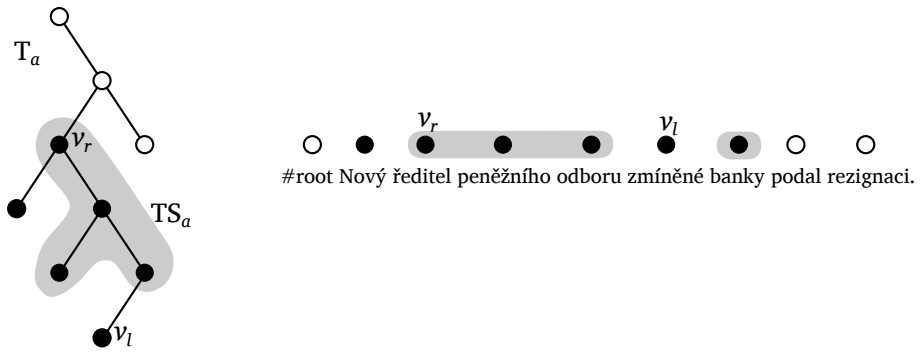
Tabulka 6.1: Počet víceslovných lexií a pojmenovaných entit uložených ve slovníku SemLex a nalezených v PDT. Celkem tedy bylo v datech nalezeno a označeno téměř 50 000 VV. Poslední sloupec ukazuje průměrný počet výskytů na jeden typ ve slovníku. Pojmenované entity do slovníku neukládáme, nemá tedy smysl mluvit o typech, rozlišujeme ovšem devět druhů (vyjmenovaných dříve), jejichž průměrná užití zde uvádíme v závorce.

Připomeňme z kapitoly 4.9, že všech téměř devět tisíc lexií SemLexu obsahuje mj. základní a lemmatizovaný tvar lexie (např. BASIC_FORM: „vyšší odborné škola“ a LEMMATIZED: „vysoký odborný škola“). Navíc jsme ke každé položce automaticky z dat vygenerovali její tektogramatickou stromovou strukturu (TREE_STRUCTURE).

[11] SemLex tak například obsahuje lexii *elektronkový počítač*, ačkoli bylo nakonec použito slovníkové heslo *elektronkový počítač 1. generace* druhého anotátora pro delší výraz v textu a uvedená kratší lexie má tedy nulovou frekvenci v datech. Podobně místo *hospodářsky vyspělá země* bylo použito *hospodářsky vyspělý stát*, místo *elektronická zařízení* (plurál), heslo *elektronické zařízení* či dokonce došlo ke změně z lexie na pojmenovanou entitu a místo *kilometry za hodinu* zůstala NE OBJECT apod.

[12] Rozlišujeme *typy*, což jsou abstraktní VV, zástupci reálných výskytů; a *instance*, což jsou realizace typů v textu. Ve slovníku tedy máme typy, v PDT jejich instance.

Tree structure (v dalším textu též zkráceně *TS*) pro víceslovný výraz VV_a je souvislý stromový graf TS_a. Uzlům tohoto stromu jsou přiřazena `t_lemma`ta. Zároveň je tento strom vždy obsažen v syntaktickém[13] stromě T_a reprezentujícím větu, ve které se VV_a vyskytuje (TS_a je podstromem[14] T_a). TS je tvořený pou-



Obrázek 6.2: Ukázka podstromu s rozvitím. Podle běžné definice by do podstromu vrcholu v_r patřily všechny plné vrcholy. Naše definice je volnější: kromě běžného podstromu tvořeného plnými vrcholy povolujeme i podstrom TS_a vrcholu v_r, který je vyznačen šedou oblastí. Plné vrcholy, které nepatří do podstromu TS_a, by pak typicky byla nějaká rozvití, která by víceslovný výraz VV_a připouštěl. List v_l (který reprezentuje řekněme přívlastek) může v povrchovém zápisu rozdělit VV na dvě části, ale stromová struktura je stále souvislá. Příkladem věty, která odpovídá tomuto stromu, může být například *„Nový ředitel peněžního odboru zmíněné banky podal rezignaci.“* — [příklad vymyšlen] s víceslovným výrazem *ředitel peněžního odboru banky.*

ze topologickou strukturou a `t_lemma`ty – součástí TS zde není ani uspořádání sourozeneckých uzlů (není potřeba), ani odkazy na pomocné uzly na a-rovině `a/aux.rf` (což zpětně považujeme za určitý nedostatek návrhu a bude vhodné je

[13] Pracujeme hlavně s tektogramatickým stromem (a tedy s *tectogrammatical tree structure, TTS*), při experimentech jsme ovšem pro srovnání také z analytických stromů získávali *analytical tree structure, ATS*. Zde tedy používáme pro oba druhy obecnější výrazy TS a syntaktický strom.

[14] *Podstromem* zde rozumíme jakýkoli stromový podgraf. Na rozdíl od definice podstromu běžné v teorii grafů tedy nevyžadujeme, aby do něj náležely *všechny* uzly od nového kořene až k listům původního stromu – stačí, když je souvislý a daný uzel je jeho kořenem. Lingvisticky řečeno, připouštíme přívlastek, který rozvíjí VV a není jeho součástí – tj. závisí na některém uzlu podstromu, ale podstrom jej neobsahuje, viz obrázek 6.2.

tam automaticky doplnit z anotovaných dat, viz část Vylepšení SemLexu v sekci 6.1.7). Z hlediska vytváření TS z původní syntaktické reprezentace celé věty lze TS nahlížet jako **úplný** souvislý stromový podgraf – tedy graf, který se skládá z podmnožiny uzlů původního stromu a ze všech hran, které mezi nimi byly; atributy uzlů jsou ovšem (s výjimkou atributu `t_lemma`) vypuštěny.

Ke každému VV v SemLexu jsme nalezli jeho výskyt v anotovaném textu. Označili jsme všechny uzly, ze kterých se VV skládá. Podstrom, který uzly tvořily, jsme uložili k heslu do slovníku.

V této sekci jsme stručně představili projekt Lexemann a jeho výstup: anotovaná data v rozsahu 800 000 slov s vyřešenými neshodami anotátorů a slovník SemLex obsahující bezmála 9 000 víceslovných lexií. Vedle anotace víceslovných lexií jsme zmínili také anotaci víceslovných pojmenovaných entit, neboť byly rovnocennou součástí projektu, nicméně my se dále budeme zabývat jen víceslovnými lexiemi, neboť ty jsou uloženy spolu s TS v SemLexu.

Nyní popíšeme způsob, jakým jsme získali další potřebná data pro experiment, který konečně popíšeme v sekcích 6.1.5 a 6.1.6.

6.1.4 Tektogramatický parsing

K přípravě automaticky syntakticky anotovaných dat pro experiment jsme použili systém Treex. Treex (Popel a Žabokrtský, 2010) je systém pro počítačové zpracování přirozeného jazyka, který obsahuje velké množství připravených modulů (*bloků*) pro jednotlivé lingvistické (či pomocné) úlohy. Každý blok má jasně specifikovaný formát vstupních a výstupních dat tak, aby bylo možné je snadno řetězit. Jejich poskládáním do tzv. *scénáře* lze zpracovat text do podoby, která je žádoucí.

Základní použití Treexu je pro překlad, ale díky vysoké modularitě je použitelný i pro řadu dalších lingvistických úkolů. Obvyklý scénář překladu sestává z desítek bloků, které postupně tokenizují, morfologicky analyzují a parsují větu (tzv. *analýza*), načež sestavený závislostní strom přeloží do závislostního stromu cílového jazyka (*transfer*). Tam se postup obrací a ze stromové reprezentace se generuje věta v cílovém jazyce (tzv. *syntéza*).

Pro naše účely postačil standardní scénář „Analysis of Czech“, jehož základní a největší bloky jsou:

- pravidlová segmentace a tokenizace,
- morfologická analýza (Hajič, 2004) a Featurama (Spousta, 2011) trénovaná na trénovací části PDT 2.0,
- MST parser s upravenými rysy (Novák a Žabokrtský, 2007) trénovaný na PDT 2.0, jehož výsledkem je a-rovina, a

- blok pro pravidlový převod a-roviny na t-rovinu.

6.1.5 Teoretické předpoklady pro vyhledávání

V předchozích sekcích jsme se seznámili s terminologií a s představou víceslovných lexií, představili jsme si, jak jsou anotovány v PDT, připomněli si slovník SemLex a předestřeli, jakým způsobem získáme automaticky a-rovinu a t-rovinu pro libovolný text. Tím jsme si připravili půdu pro popis samotného experimentu identifikace VV v textu.[15] Ačkoli je naše úloha řešena čistě pravidlovou metodou, budeme pro data používat pojmy známé ze statistických metod:

trénovací data pro ručně anotované texty, v nichž jsou VV vyznačeny ručně a z nichž čerpáme informace, a

testovací data pro zpracovávaný úsek textu, kde jsou VV vyhledávány automaticky (a je na nich případně měřena úspěšnost).

Mějme tedy jak trénovací tak testovací data doplněna o informačně bohatou tektogramatickou rovinu. Jak v její reprezentaci vypadá nějaké slovní spojení (při pohledu na všechny možné tvary a výskyty v jazyce)?

- Díky morfologické analýze a lemmatizaci splynou různě skloňované, či jinak ohýbané výrazy. ⇒ Získáme `t_lemma`.
- Díky syntaktické analýze a zanedbání pořadí uzlů se stírají rozdíly způsobené různým slovosledem a některými dalšími druhy obměn (např. vkládáním). ⇒ Získáme stromovou strukturu.
- Díky tektogramatické rovině zmizí pomocná slova, například modální a pomocná slovesa, a doplní se některá povrchově vypuštěná slova. ⇒ Stromovou strukturu zjednodušíme a zobecníme.

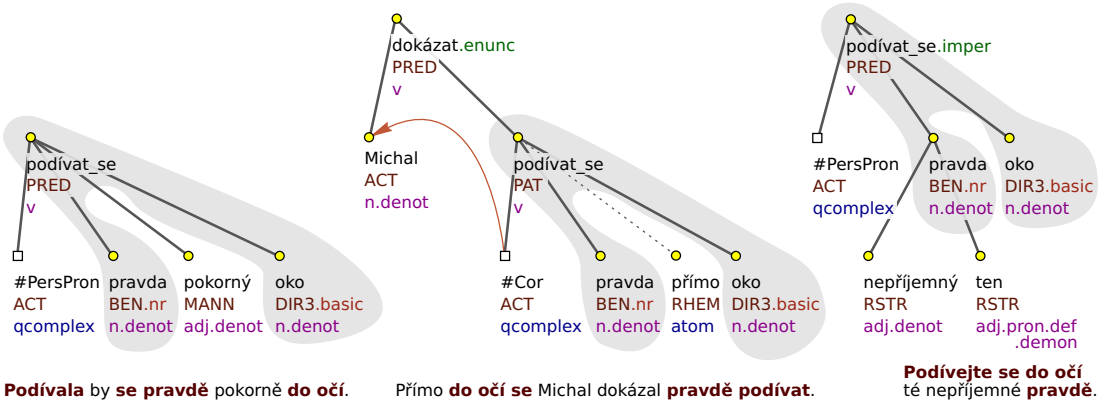
Jak to může vypadat vidíme na obrázku 6.3. Ač se VV různě obměňuje, ty vlastnosti, které na něm sledujeme, zůstávají stále stejné.

Vezměme tedy vše, co o nějaké víceslovné jednotce W ve slovníku SemLex máme, a hledejme stejná místa jinde ve stromech. Předpoklad je, že najdeme právě všechny výskyty W. Tedy že můžeme postupovat následovně:

> Pokud je část věty označena jako víceslovný výraz W v trénovacích datech, označíme jako W každý *shodný* úsek v testovacích datech.

Myšlenka, která stojí za naší metodou, je v zásadě jednoduchá. Stojí však na komplexní teorii FGP a spoléhá na bohaté informace, které poskytuje anota-

[15] Ten je popsaný v článku Bejček et al., 2013, jehož odpovídající část tato sekce výrazně rozšiřuje.



Obrázek 6.3: Víceslovný výraz *„podívat se pravdě do očí“* v různých užitích ve třech větách. Ačkoli povrchová realizace se značně liší, jejich reprezentace na t-rovině je shodná. (Věty jsou záměrně zkonstruovány tak, aby se VV co nejvíce měnil, možná až na hranici únosnosti. Přesto i takovým větám, vyskytnou-li se, přísluší stále stejná tektogramatická reprezentace.)

ce z FGP vycházející. Máme-li takto kvalitní data[16] s anotací VV, prohledávací procedura je spíše technická záležitost. (Obtíže se však ještě skrývají ve slově *„shodný“* ve větě uvedené výše. Co znamená „shodný“ a jaké změny mohou nastávat rozebereme v rámci hypotézy B na následujících stranách.)

K dispozici tedy máme (viz sekce 6.1.3) přes 700 000 slov ve větách s doplněnou tektogramatickou reprezentací, kde jsou ručně vyznačeny všechny víceslovné výrazy. Je pak snadné přiřadit ke každému typu VV z anotovaných dat jeho tektogramatickou reprezentaci (tedy vytvořit slovník VV s přiřazenými TS). A opačně je zase snadné vyhledat všechny podstromy *shodné* s tektogramatickou reprezentací zvoleného VV v tektogramaticky anotovaných datech (tedy identifikovat VV v textu).

Bude to však stačit? Najdeme tak všechny VV v textu? Nemůžeme chybně najít něco navíc? Jak dobrou tektogramatickou reprezentaci dokážeme vytvořit bez ruční práce? Co dělat v případě více možných reprezentací pro jeden VV? A neukáže se nakonec jako úspěšnější přece jen řešení, které je jednodušší a pracuje pouze s morfologicky označkovanou větou, příp. větou zpracovanou na a-rovině?

[16] V případě automaticky vytvořené t-roviny je samozřejmě kvalita nižší. Navíc na následujících stránkách ukážeme, že ani ručně vytvořená t-rovina neodpovídá tektogramatické rovině FGP do té míry, do jaké bychom chtěli.

Vlastnosti reprezentace VV pomocí TS

Následující výčet vlastností TS do značné míry určuje odpovědi na položené otázky. Tyto vlastnosti jsme před začátkem manuálních anotací, nebo v jejich průběhu, zformulovali jako hypotézy, které bylo potřeba ověřit. Každou vlastnost popíšeme, vysvětlíme, k čemu je její platnost dobrá, a uvedeme, do jaké míry se potvrdila.

Předpokládali jsme tyto vlastnosti TS na tektogramatické rovině:

A) **Souvislý podstrom:** Každá víceslovná lexie je vždy reprezentována *souvislým* podstromem (str. 146).
B) **Jediná TS pro lexii:** Jedna položka ve slovníku má *jedinou* TS (str. 148).
C) **Nikdy doslovně:** TS ze slovníku nalezená v datech představuje *vždy* VV, nikoli doslovný význam (str. 156).
D) **Unikátnost TS:** Jedna TS se ve slovníku vyskytuje jen jedenkrát, neopakuje se (str. 158).

První dvě hypotézy uvádí již Straňák (2010, od strany 55).[17]

Souvislý podstrom (hypotéza A). Příkladem souvislého podstromu je obrázek 6.4a a 6.4b, příkladem „nesouvislého podstromu“ obrázek 6.4c.

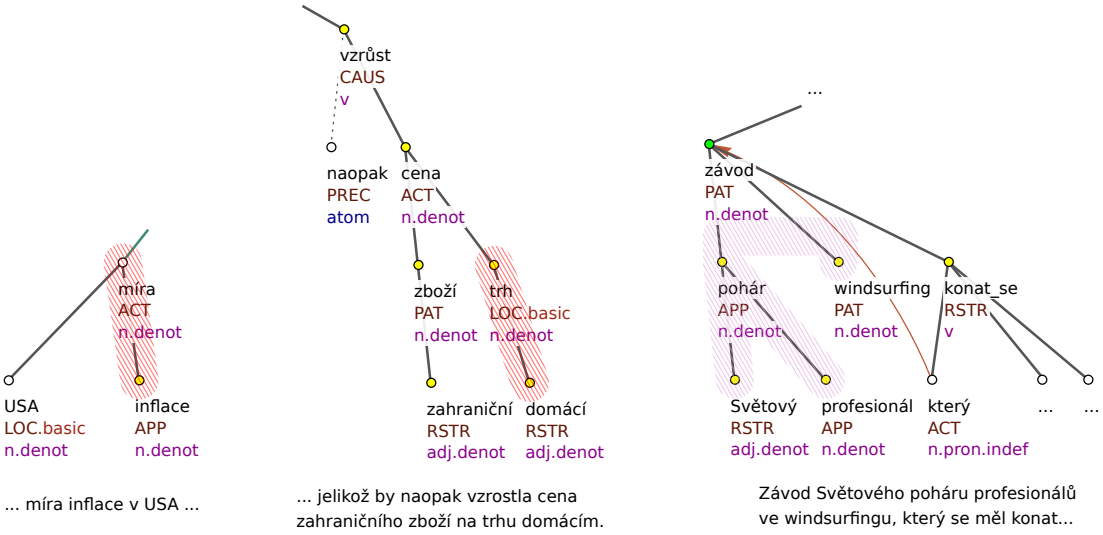
Budou-li všechny víceslovné lexie reprezentovány souvislým podstromem, budeme mít pro každou položku v SemLexu přehlednou kompaktní reprezentaci.

Tato vlastnost není sice úplně zásadní, ale její platnost velmi usnadní vyhledávání TS v testovacích datech: je-li přesně určené, v jakém syntaktickém vztahu mezi sebou hledané uzly mají být, snadněji vyloučíme jejich přítomnost v datech. Když povolíme nesouvislost, přibývají otázky, jak moc se od sebe mohou jednotlivé komponenty souvislosti ve stromě vzdálit a v jakém vztahu mohou ještě být, aby celek tvořil VV. K jejich zodpovězení lze jistě opět použít data – pokud jich ovšem bude dostatek k tomu, abychom měli jistotu, že jsme viděli všechny povolené varianty.

Hypotéza je důležitá i z hlediska teoretické reprezentace VV, tedy i odhlédnemli od dosud probíraného praktického hlediska. V PDT od verze 2.5 je celý VV reprezentován jako jediný t-uzel (viz sekce 3.4), což vychází z představy FGP o členech tektogramatické roviny. Neplatnost této hypotézy by zpochybnila oprávněnost takové reprezentace pro některé VV, nebo by přinejmenším zpochybnila označení takových slovních spojení za víceslovné výrazy.

Hypotéza platí zhruba pro 98,8 % typů VV (99,6 % výskytů v datech). Výjimkou je technické řešení u cizích slov (všechny členy cizojazyčné fráze jsou

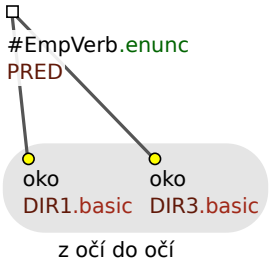
[17] Hypotézu A potvrzuje s výjimkou koordinačních struktur. Pro hypotézu B uvádí protipříklady (zde budou uvedeny pod čísly 2, 5 a 6). K řešením, která pro ně navrhuje, se níže přikláníme i my.



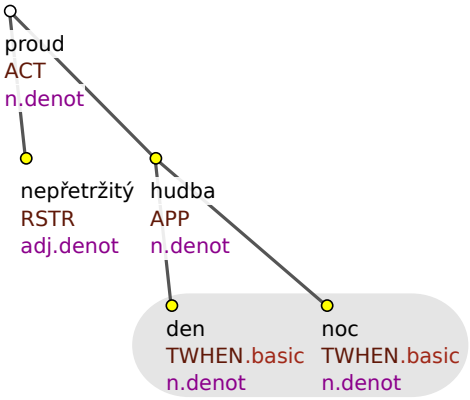
Obrázek 6.4: Ukázky VV reprezentovaných souvislým a nesouvislým grafem. První dva stromy obsahují víceslovné lexie (*míra inflace* a *domácí trh*), které jsou reprezentovány souvislým podstromem.
Pokud bychom chtěli označit například *inflace v USA*, nebo *zboží na trhu*, což jsou sice úseky věty, ale syntakticky zcela chybně vytržené (jde o *cenu zboží* a *cenu na trhu*), výsledek by byl nesouvislý, což je zcela v pořádku.
Třetí strom naopak obsahuje nesouvislý podstrom pro VV (v tomto případě spíše pojmenovanou entitu typu objekt než lexii) *Světový pohár profesionálů ve windsurfingu*. Zde se očividně neshodl anotátor PDT, který *windsurfing* z názvu poháru vyňal, s anotátorem VV, který se domníval, že je to součást názvu.

anotovány uzly závislé na technickém uzlu #Forn) nebo některá ustrnulá spojení jako např. *„z očí do očí“*, *„ve dne v noci“*, *„jakýs takýs“*, nebo *„otevřená zadní vrátka“*, kde všechny části závisí na (často nevyjádřeném) slovese, které už do frazému nepatří – viz příklady na obrázcích 6.5 až 6.8. Některé z nich měly být zřejmě už při anotaci PDT 2.0 řešeny frazeologicky. Obrat *z očí do očí* rozhodně nevyjadřuje pohyb odněkud (DIR1) někam (DIR3), ale společně vyjadřuje nějaký způsob jednání, navrhujeme proto uzel s `t_lemma`tem *oko*$_{\text{MANN}}$ a závislý uzel *do_očí*$_{\text{DPHR}}$. Takto je již podstrom VV souvislý. Podobně *den*$_{\text{TWHEN}}$ *v_noci*$_{\text{DPHR}}$ a *takový*$_{\text{RSTR}}$ *jakýs*$_{\text{DPHR}}$.

[18] Anotace PDT 2.0 zde vyžaduje sloveso do té míry, že když ve větě přítomno není (fráze je použita samostatně), musí zde být doplněn uzel pro „empty verb“ jako na obrázku.



Obrázek 6.5: Fráze *„z očí do očí“* v PDT 2.0 vyžaduje jako svou součást sloveso.[18] Toto sloveso se ovšem může lišit větu od věty: *„Neuspěl ani Bielik z očí do očí proti sparťanu Koubovi.“* — [PDT 2.0, zkráceno], *„Promluvil si s ním z očí do očí.“*, *„Vyříkají si to z očí do očí.“*. Proto ho za součást VV nepovažujeme a z obou těchto důvodů (anotace t-roviny i naše chápání VV) vznikne nesouvislá TS.



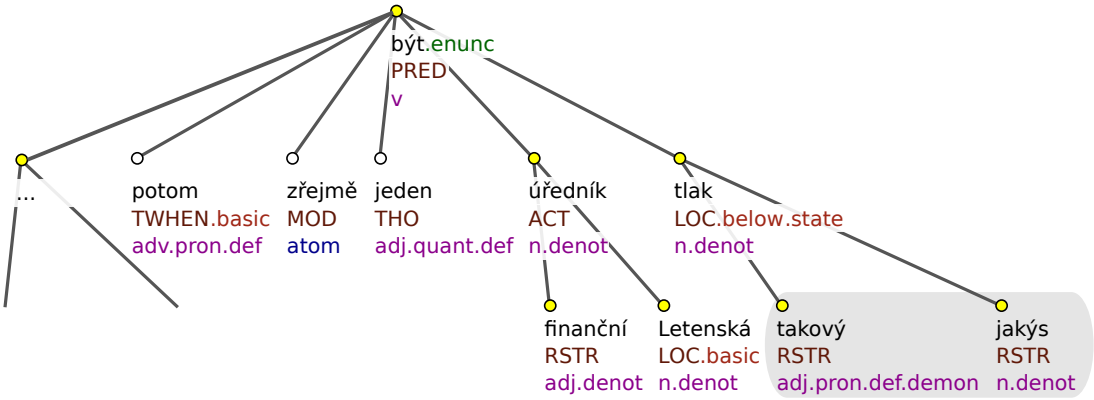
Obrázek 6.6: Fráze *„ve dne v noci“* v PDT 2.0 je rozdělena, anotace ji nechápe frazeologicky, ale pouze jako dvě samostatná příslovečná určení času. Přes zřejmou frazeologickou povahu nejsou dva t-uzly spojeny spolu, nýbrž s řídícím uzlem.

Za výjimku z tohoto pravidla nepovažujeme rozličné případy spojené s koordinací. Jejich souvislost či nesouvislost totiž závisí jen na tom, jak budeme pracovat s relací rodiče a efektivního rodiče. VV „nesouvislé“ kvůli koordinaci lze ve slovníku zachytit jako souvislé, ať už jde o výrazy částečně koordinované (jako kytovec *vorvaňovec pacifický* ve spojení *vorvaňovec pacifický i severomořský*), či koordinaci přímo obsahující (jako *alfa a omega*).

Poznamenejme závěrem, že souvislý podstrom můžeme vyžadovat mj. proto, že pracujeme se závislostními stromy. Jak vypadají dva víceslovné výrazy ve složkové anotaci Penn Treebanku vidíme na obrázku 6.9 z PCEDT.

Jediná TS pro lexii (hypotéza B). Tato vlastnost by zaručila kladnou odpověď na otázku „Najdeme všechny VV v testovacích datech?“.

Znamenalo by to, že by stačilo daný VV vidět jen jedenkrát v trénovacích datech a byla by jistota, že ho (ve správně syntakticky anotovaných testovacích datech) nikdy nepřehlédneme; tedy zaručeně žádné případy „false negatives“ pro VV ze slovníku. (Stále by tu ale bylo riziko „false positives“, tedy že daná TS se může vyskytnout, aniž by reprezentovala VV, viz vlastnost C.)

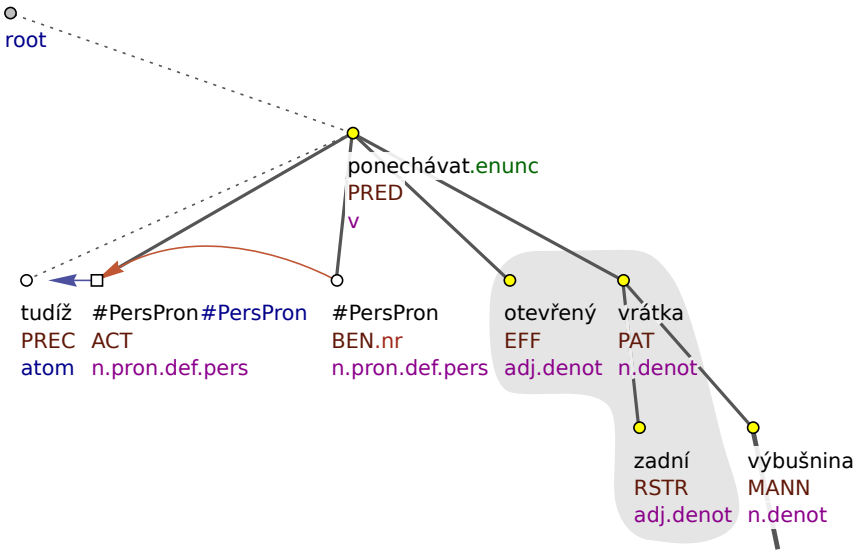


Obrázek 6.7: Fráze *„jakýs takýs“* v PDT 2.0 (lemmatizovaná jako *takový* a *jakýs*) představuje stejný problém jako *z očí do očí* (6.5) a *ve dne v noci* (6.6).

Tato hypotéza se ovšem nepotvrdila. Stručně řečeno, platí pro „učebnicové VV“ a platila by (s výjimkou 9, 10 a 11) za pozměněných anotačních pokynů pro tvorbu t-roviny. Se současnými pravidly t-roviny však pro nejrůzněji modifikované VV z běžného textu často neplatí.

Jeden VV může mít více TS v případech vyjmenovaných a vysvětlených na následujících stránkách. Jde namátkou o zkratky, elipsy, synonyma, zdrobněliny, vidové dvojice apod. Všechny tyto rozličné typy mají společné to, že dvojice výrazů zde odkazují ke stejné (nebo podobné) entitě, konceptu ve skutečném světě pomocí různých slovních vyjádření. V důsledku odlišných slov v povrchovém vyjádření se pak liší i `t_lemma`ta v hloubkové reprezentaci (případně některá slova a tudíž i jim odpovídající t-uzly chybí). Přesto jsme jim přiřadili stejnou položku slovníku, protože považujeme za čistší řešení mít jednu položku pro jeden koncept, nehledě na konkrétní vyjádření.

Seznam jevů je zhruba uspořádán od nejzřejmějších, v zásadě synonymních dvojic (1–4), přes složitější případy, které by pro shodnou reprezentaci vyžadovaly zásah do t-roviny PDT (5–8), až k diskutabilním dvojicím, které spolu souvisí méně (9–11). Dalo by se tedy namítat, že některé z dvojic na konci seznamu *nemají* dostatečně podobný význam a neměli bychom jim tudíž přiřazovat tutéž slovníkovou položku. Hranice mezi „dostatečně podobnými“ VV (sdílejícími stejné heslo ve slovníku) a „významně rozdílnými“ VV (které tak dostanou dvě odlišná hesla) rozhodně není ostrá. Je do značné míry věcí našeho rozhodnutí, které VV budou sdílet slovníkovou položku. Dá se říci, že jsme vědomě stanovili



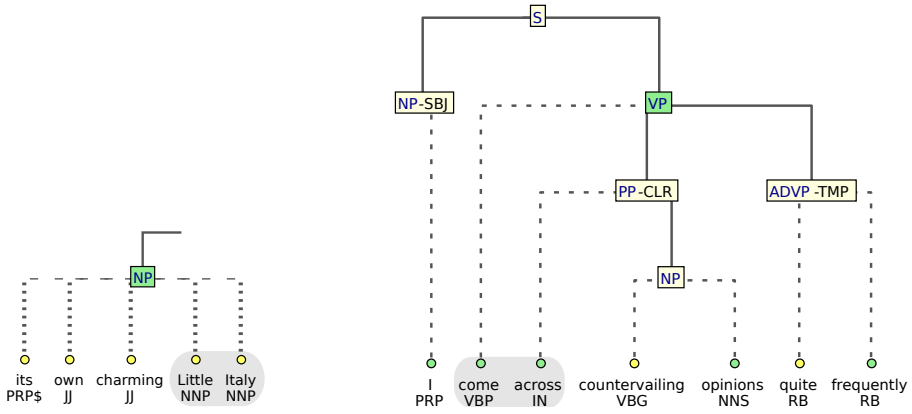
Obrázek 6.8: Ve frázi *„otevřená zadní vrátka“* není *„otevřená“* považováno za atribut RSTR vrátek, nýbrž za EFF slovesa (které z našeho pohledu opět do fráze nepatří, neboť se může měnit: *mít, nechat, připravit si, počítat s, spoléhat se na, …*). Z jediného výskytu v trénovacích datech však nemůžeme usoudit, zda s jiným slovesem nemůže být fráze jinak strukturovaná, například souvislá.

tuto hranici podobných VV tak „přísně“, že vlastně požadujeme, aby prvních osm z následujících jazykových jevů (1–8) pravidlo jediné TS pro VV porušovalo. Považujeme to za lingvisticky adekvátnější popis.

Z technického hlediska je možné požadavek jediné TS snadno obejít: lze si uložit do slovníku všechny TS a hledat jen nejčastější TS, nebo naopak hledat všechny.

Je to však řešení jen zdánlivě. Představme si VV, který je známý (= je uložený ve slovníku včetně svých TS, které se v trénovacích datech vyskytly) a tedy očekáváme, že při automatickém vyhledávání ho vždy nalezneme. Pokud ovšem hypotéza jediné TS pro VV nebude platit, může existovat jiná varianta TS, kterou jsme dosud neviděli, ve slovníku ji nemáme, a proto ji najít nemůžeme. Nefunguje tedy jakási generalizace případů.

Jakákoli technická řešení (zmíněné ukládání všech možných TS k jednomu heslu, rozdělení VV do více slovníkových hesel, zvětšení trénovacích dat) jsou tedy jen částečná, neboť je možnost generalizace stále omezena a najdeme jen ty varianty TS, které jsme už viděli.



Obrázek 6.9: Víceslovný název označující charakter městské čtvrti – *Little Italy* – a frázové sloveso *come across* v Penn Treebanku. Ani jednomu VV nepřísluší taková část složkového stromu, která by tvořila souvislý podstrom.

Proto pro adekvátní popis preferujeme změnu t-roviny tak, aby některé dvojice splynuly a v maximálním počtu případů byla zaručena jediná TS pro víceslovný výraz. Jsme si ale vědomi faktu, že všechny níže vyjmenované typy takto vyřešit nepůjdou.

Tektogramatická rovina (a potažmo její odraz v t-rovině PDT) je rovinou jazykového významu. Každý uzel stromu by měl nést samostatný význam. Proto je ostatně od PDT 2.5 (viz sekce 3.4) každý VV zobrazen jako jediný uzel. Uzel má nést tři informace: funktor, gramatém a *sémantém*. Sémantém vyjadřuje význam a má být realizovaný jako odkaz do slovníku – více synonymních uzlů by mělo podle FGP nést stejný odkaz; v PDT, kde tomu zhruba odpovídá `t_lemma`, se ovšem takto daleko nešlo ani ve verzi 3.0.

U první skupiny případů je potřeba společného zachycení mnohem zřetelnější (body 1–4). Jedná se o synonymní[19] víceslovné výrazy (přičemž synonymie je dosahováno čtyřmi různými prostředky). Důležité přitom je, že většina z těchto prostředků (například zkratka či elipsa) může být použita kdykoli a pro jakýkoli

[19] V této práci chápeme výraz *synonymní* jako „reprezentovaný jedním heslem ve slovníku“ a potažmo tedy „odkazující ke stejné skutečnosti“ jako například v článku o *„Českém svazu plaveckých sportů“* je pro něj synonymní *„plavecký svaz“*. Nebo v kontextu, kde *„OSN“* znamená *„Organizace spojených národů“* (a nemluví se třeba o *„Ottově slovníku naučném“*), je tato zkratka synonymní s *„Organizací spojených národů“* – a také s *„UN“*, chce-li někdo použít anglickou zkratku. Další příklady: *„cenová hladina“* ~ *„hladina cen“*; *„držet v tajnosti“* ~ *„udržovat v tajnosti“*, …

VV. Není proto praktické pro tyto varianty VV zavádět nová slovníková hesla, ačkoli by bylo možné je mezi sebou propojit relací synonymie.

1) zkratka Část víceslovného výrazu (případně i výraz celý[20]) může být zkrácena (nastává to zejména pro názvy institucí a států a jejich zkratky velkými písmeny). Tato zkratka má stejný význam jako celý výraz a měla by tudíž na tektogramatické rovině mít jako celek stejné `t_lemma`. (Přiřazení stejného `t_lemma`tu ovšem často vyžaduje sloučení více uzlů, což předpokládá už vyřešenou anotaci alespoň těchto VV.)
↪ *„druhá vlna kuponové privatizace / druhá vlna KP“*, *„země bývalého Sovětského svazu / země bývalého SSSR“* (zde ještě nastává záměna na *Sovětský svaz* namísto *Svazu sovětských socialistických republik*.), *„Václav Havel / V. Havel“*, *„Rada bezpečnosti OSN / RB OSN“*, …

2) elipsa Elipsa je nejtěžším jevem, se kterým se musíme vypořádat. Víc různých víceslovných výrazů, které se však liší pouze vypuštěním slov, může odkazovat k jedné a téže skutečnosti. K elipse ve VV dochází například ze stylistických důvodů, je-li výrazu opakovaně používáno ve stejném textu. Můžeme rozlišit *gramatickou elipsu* (např. *„mlčeti zlato“*, zde je – při každém použití tohoto VV – elidováno nějaké sloveso, např. si můžeme doplnit sloveso *je*) a *aktuální elipsu* (např. *„liga“*, když je z kontextu jasné, že se jedná o *fotbalovou ligu*).
Gramatická elipsa je vždy stejná – slova budou z výrazu vynechaná stejně tak ve větách, se kterými pracovali anotátoři, jako ve větách, kde budeme hledat VV automaticky. Neprojevuje se tudíž porušením vlastnosti B (jediná TS pro lexii).[21]
Pokud ovšem víme, jaké slovo bylo vypuštěno (ať už se dá doplnit z okolního kontextu, nebo ze znalosti světa), jedná se o aktuální elipsu – a tou se zabývat musíme. Zvlášť v kombinaci se zkratkou (předchozí pasáž) může jeden VV nabývat velkého počtu variant, pro které chceme mít uložené jedno společné slovníkové heslo.[22] Některé aktuální elipsy řeší tektogramatický

[20] V takovém případě již ovšem nejde o VV (z hlediska počtu slov či počtu t-uzlů) a v projektu popisovaném v části 6.1.3 nebyl anotován.

[21] Přesto musíme s gramatickou elipsou uvnitř VV pracovat. Výrazy jako *„mlčeti zlato“* jsou na t-rovině zachyceny s pomocí umělého slovesného uzlu #EmpVerb za vynechaný sponový predikát a umělého uzlu #Gen pro všeobecného konatele mlčení. Oba tyto umělé uzly (které jsou přítomny v tektogramatické analýze každého výskytu *mlčeti zlato* a tím tento výraz spoluurčují) zahrnujeme do syntaktické reprezentace VV. Díky tomu také zachováme vlastnost A.

[22] Alternativně bychom mohli uchovávat eliptické a zkrácené varianty ve slovníku zvlášť. Pokud připustíme, že mají totožný význam jako plné tvary VV, a přidáme k nim odkaz na

manuál (Mikulová et al., 2006, str. 417) doplněním elidovaných uzlů, jde ovšem pouze o souřadná spojení a některé konstrukce s kontrolou. Zbylá elidovaná slova na t-rovině zastoupena nijak nejsou a pro hledání VV jsou tedy problémem.

↪ *„druhá vlna kuponové privatizace / druhá vlna privatizace / druhá vlna“ „neviditelná ruka trhu / neviditelná ruka“* ve větě *„Trh není schopen řešit svou ‚neviditelnou rukou‘ všechny problémy.“*, *„Český svaz plaveckých sportů / Český svaz / plavecký svaz“*, *„sirup / škrobový sirup“* (z textu je zřejmé, že se stále jedná o *škrobový sirup*), *„ministerstvo práce a sociálních věcí / ministerstvo práce“*,…

3) synonymum Naproti tomu synonyma jsou problémem poměrně malým. Jde o případy, kdy část VV nahradíme jejím synonymem bez porušení celkového významu VV.[23] Opět pro oba VV zavádíme jedinou položku slovníku. Původní a nový VV se dokonce mohou lišit počtem slov. Mezi touto kategorií synonym a lexikální variantou (bod 9 zmíněný níže) nevede ostrá hranice. Dvojice v této kategorii by měla být synonymní vždy, zatímco při lexikální variantě hodně záleží na kontextu, do něhož je VV zasazen.

↪ *„dopingový test/zkouška“*, *„země/stát bývalého Sovětského svazu“*, *„pohraniční/příhraniční oblast“*, *„koncentrační tábor / koncentrák“*, *„osoba/ /pracovník/občan/zaměstnanec se změněnou pracovní schopností“* (zde je možné si povšimnout, že ačkoli jednotlivá slova nejsou obvykle svými synonymy, v kontextu této VV jsou synonymní vždy: kontext *pracovní schopnosti* vyvolává koncept *zaměstnání* a ze všech lidí (*osob, občanů*) se stávají *zaměstnanci* a *pracovníci*), *„státní zástupkyně/zastupitelka“*, *„2./druhá vlna kuponové privatizace“*, *„Stanleyův/Stanleyho pohár“*, *„surové dříví/dřevo“*, *„stavebněmontážní/stavebně montážní společnost“* (opět případ, kdy se liší i počet slov),…

4) přívlastek (ne)shodný Ke změně z přívlastku shodného (adjektivum před rozvíjeným jménem) na neshodný (substantivum za jménem), či naopak dochází často také bez změny významu.

↪ *„druhá vlna privatizace / druhá privatizační vlna“*, *„cenová regulace / regulace cen“*, *„premonstrátský řád / řád premonstrátů“*,…

synonymní VV, dospějeme k provázané skupince slovníkových hesel, v níž všechna hesla mají zcela shodné hodnoty atributů, jen TS mají odlišnou. Což je vlastně jen méně úsporná a méně přehledná, ale obsahově shodná varianta k řešení, kdy máme jednu slovníkovou položku a k ní si poznamenáme větší počet variant TS, pokud se vyskytují.

[23] Takové víceslovné lexie sice umožňují substituovatelnost (jedna z vlastností na straně 136), což na VV nenapovídá, ale podle zbylých kritérií jsou stále za lexie považovány.

Druhá skupina neobsahuje synonyma, vykazuje vždy mírnou odchylku významu. Ve všech případech (zdrobněliny, přechylování, vid a slovní druh; body 5–8) jsou si ale dvojice významem blízké do té míry, že jsme se rozhodli ukládat je do slovníku vždy společně. Hlavním důvodem bylo, že se domníváme, že na tektogramatické rovině (rovině jazykového významu) mají mít shodnou lexikální reprezentaci (stále se liší gramatémem). Chtěli bychom však zdůraznit, že kdybychom se rozhodli je společně neukládat (nebo kdyby už nyní byla na t-rovině `t_lemma`ta uvedených dvojic[24] sjednocena), většina zmíněných porušení vlastnosti B by nenastávala.

5) zdrobnělina Dvojice VV s téměř totožným významem a se stejnou idiosynkracií (např. tvorbou, ustrnulostí, stejně se vymykající jazykové pravidelnosti apod.), které se liší zdrobněním jednoho slova. Navrhujeme pro PDT zavedení nového gramatému, o kterém FGP nehovoří a který by zdrobnělé (a podobně zveličelé) názvy odlišoval; `t_lemma` by potom pro ně bylo společné, což navrhuje i Straňák (2010).
↪ *„mateřská škola / mateřská školka“* (ačkoli jedno slovo je zdrobnělé, na významu VV se nezmění vůbec nic: nemáme větší a menší mateřské školy), *„soudek s prachem / sud s prachem“* (kdyby oba VV stály v kontrastu, dalo by se možná říci, že autor chce naznačit, že *„sud s prachem“* skrývá větší výbušný potenciál; běžně jsou to však synonyma), *„chovný pár/párek“*, *„rodinný dům/domek“*, …

6) přechylování Dvojice VV se liší pouze tím, zda je v mužském nebo ženském rodě. Domníváme se ve shodě se Straňákem (2010), že na t-rovině by měly být zachyceny stejným `t_lemma`tem. Lišit by se měl podle nás jen gramatém jmenného rodu `gender`. (Ten rovněž nemá žádný předobraz v FGP.) Může to být užitečné i prakticky: V nějaké fázi je potřeba poznat, že učitel a učitelka jsou tentýž pojem lišící se pouze rodem. Takový slovník mužských a ženských ekvivalentů patří spíše do derivační morfologie – tedy do tektogramatického parsingu – než do rozpoznávání VV. Bude to univerzálnější řešení než zpracování té podmnožiny případů vyskytujících se ve VV.
↪ *„mistr/mistryně světa“*, *„státní zástupce/zástupkyně“*, *„poštovní doručovatelka/doručovatel“*, *„stálý dopisovatel / stálá dopisovatelka“*, …

7) slovesný vid Víceslovný výraz obsahuje sloveso (nebo deverbativní substantivum či adjektivum), které má jak dokonavou, tak nedokonavou variantu. Slova lišící se jen videm by podle teorie FGP neměla mít na tektogramatické

[24] Navrhujeme, aby se tak stalo pro všechny body 5–8: zdrobněliny, přechylování, slovesný vid i slovní druh.

rovině různá lexikální vyjádření – v současném PDT však na t-rovině mají různá `t_lemma`ta. Již nyní je u sloves přítomen gramatém slovesného vidu, má ale pouze informační nikoli rozlišovací charakter: vidové protějšky mají sice různý gramatém `aspect`, ale stejně tak mají různá `t_lemma`ta, takže je tatáž informace uložena duplicitně. Gramatém by měl být zachován, ale vidové protějšky by měly být reprezentovány stejným `t_lemma`tem.
↪ *„pohlavně zneužít/zneužívat“*, *„zneužití/zneužívání pravomoci veřejného činitele“*, *„odvolací/odvolávací řízení“*, *„dostat/dostávat se na tenký led“*, …

8) slovní druh Některé VV je možné modifikovat do překvapivé míry – například změnit slovní druh některých slov, ze kterých se výraz skládá. V takovém případě se podstata významu VV nezmění, jen se, řekněme, z věci stane její vlastnost nebo z děje jeho důsledek. Stále si však zachovává stejné vlastnosti VV. Navrhujeme využít informace o slovním druhu (na jeho základě by bylo při syntéze textu možné z *dělat* vytvořit *dělání*) a `t_lemma` mít stejné pro příbuzná derivovaná slova.
↪ *„působit dojmem / působící dojmem“* (sloveso se mění na adjektivum a stejně tak se mění i slovní druh celého[25] VV), *„sociální demokracie / sociálně demokratická“* (adjektivum se substantivem se mění na adverbium s adjektivem a slovní druh celého výrazu se mění ze substantiva na adjektivum), *„právo nakládat/nakládání s bytem“* (v tomto případě se dokonce význam ani slovní druh celé VV záměnou substantiva za sloveso vůbec nezmění), …

Třetí skupina je diskutabilní, nicméně anotátoři i takovéto dvojice občas slučovali. Ačkoli je zřejmé, že následující dvojice VV odkazují (obvykle) k jiné skutečnosti, přesto spolu zřetelně sdílí některé vlastnosti víceslovného výrazu, způsob tvoření, posun významu apod.

9) lexikální varianty Existují VV, které umožňují na jedné své pozici obměňovat slova. Obvykle jsou možnosti, ze kterých lze vybírat, velice omezené (sémanticky, nebo i výčtem). Někdy se varianty liší tak málo, že jsou (alespoň v některém kontextu) nahraditelné.
↪ *„být/končit/uváznout na mrtvém bodě“*, *„být/žít v klidu“*, *„být/hrát na špici“*, …

[25] Víceslovné výrazy vystupují jako jednotky jazykového popisu – leckdy je možné je parafrázovat jedním slovem – a plní v něm jako celek nějakou syntaktickou funkci. Má proto dobrý smysl mluvit o *slovním druhu celého VV*. Uvedený příklad lze parafrázovat jako *„vypadat (jako) / předstírat“*, resp. *„vypadající (jako) / předstírající“*, tedy jako sloveso, resp. adjektivum.

reflexivita je zvláštním případem lexikálních variant. Mluvíme o takových VV, které existují ve dvou variantách: jednou je jejich součástí reflexivní forma slovesa, podruhé nikoli. Jedná se obvykle o *derivovaná reflexiva tantum*,[26] tedy o spontánní, nevědomou činnost a morfém *se/si* nelze nahradit reflexivním zájmenem *sebe/sobě* (*uhodit sebe* ≠ *uhodit se* ~ *uhodit sebe omylem*). Ačkoli se tím změní osoba, na níž je činnost (záměrně či samovolně) konána, tvoření významu víceslovného výrazu se tím nezmění a oba tvary tedy chceme mít sdruženy v jedné slovníkové položce. V případě samovolné činnosti jde ovšem o jiný význam a jiné `t_lemma`, neboť k němu ono *se/si* patří.
↪ *„změnit k lepšímu / změnit se k lepšímu“* (jak ukazuje obrázek 6.10, tektogramatické reprezentace se v těchto případech liší), *„vzít život / vzít si život“* (zde opět ve zvratné variantě vystupuje `t_lemma` *vzít_si*), …[27]

10) hyperonymum Dvojice VV, která zařazujeme do této skupiny, nahrazují jedno ze svých slov slovem nadřazeným/podřazeným, obvykle dokonce příbuzným (lišícím se příponou/předponou). Význam VV se sice mění, ale cítíme významné propojení významu a shodné vlastnosti a tvorbu výrazu.
↪ *„občanský zákon/zákoník“* (možná dokonce patří k synonymům), *„informační dálnice/superdálnice“*, *„kopírovací stroj/přístroj“*, *„důchodové pojištění/připojištění“*,…

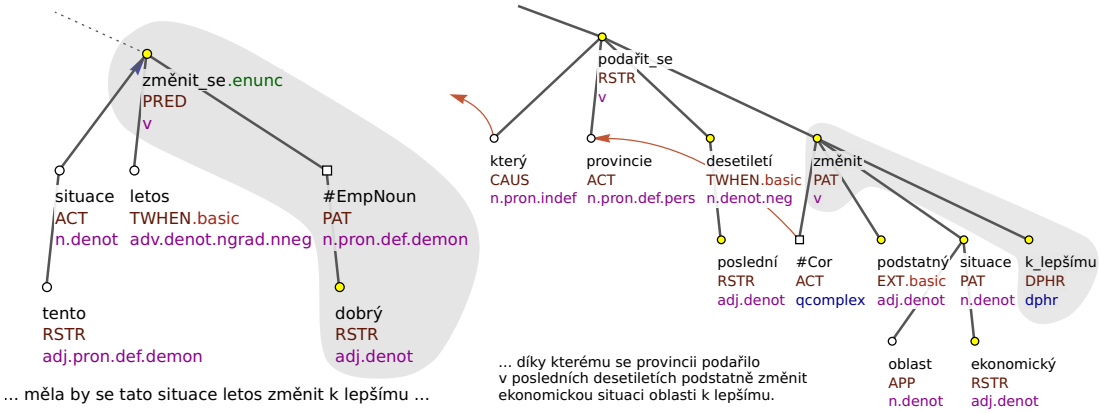
11) meronymum Zde je jedno slovo pro jednotlivost nahrazeno slovem pro celek. Význam se liší výrazně, neboť již jde o kvalitativně jinou věc, ale jazykové vlastnosti a tvorba je stále pro obě VV stejná.
↪ *„sociální demokracie/demokraté“*, *„městská policie / městský policista“*, *„policejní ředitelství/ředitel“*

Pozorný čtenář si jistě povšiml, že některé VV podstupovaly více typů variování. Například *„druhá vlna kuponové privatizace“* má dvě varianty pro *druhá/2.*, dvě ortografické varianty pro *kuponová*, dvě možnosti elipsy, jednu zkratku a navíc záměnu za adjektivní formu shodného přívlastku, takže může vytvořit až $2 \times (2 + 2 + 1 + 1) = 12$ možností (ne všechny jsou naštěstí obsaženy v datech).

Nikdy doslovně (hypotéza C). Tady si klademe otázku, zda v datech každý výskyt TS shodné s tou ve slovníku představuje nutně víceslovný výraz. Zejména tedy zda se nemůže jednat o doslovné užití daného slovního spojení.

[26] Na rozdíl od reflexiv, která vyjadřují děj záměrně vztažený na sebe (*umýt se* = *umýt sebe*, *koupit si* = *koupit sobě*), nepopisují *derivovaná reflexiva tantum* vědomou činnost aktora děje.

[27] Jiný případ je *„dávat si pozor / dávat pozor“*, kde je možné zvratné *si* vypustit bez změny významu (a ovšem neplatí to pro *dávej (si) na něj pozor*).



Obrázek 6.10: Tektogramatická reprezentace fráze *„změnit se k lepšímu“* (vlevo) a *„změnit (něco) k lepšímu“* (vpravo). Vidíme, že reflexivní varianta se v anotaci PDT výrazně liší – má jiné `t_lemma` i jinou strukturu.

- Odlišnost v `t_lemma`tu (viz též poznámka 26 na str. 156) je v pořádku – mimoděčné *změnit se* je chápáno jako jiný lexém a tudíž i jiný význam než *změnit* (kam patří i význam *změnit sebe*), což je záměrný děj.
- Odlišná struktura je s největší pravděpodobností chyba anotace PDT 2.0, protože PDT-Vallex opomněl sjednotit valenční rámce reflexivní a nereflexivní varianty slovesa *změnit* pro idiomatická spojení s výrazem *k lepšímu* – v obou případech by měl existovat samostatný rámec, v němž výraz *k lepšímu* bude ohodnocen funktorem pro idiom, tj. DPHR. Nutno však dodat, že takováto sjednocení není možné provádět pro všechny reflexivní-nereflexivní dvojice a výsledek je často sporný.

Shrňme, že v PDT 2.0 to zkrátka takto odlišně zachyceno bylo, tudíž pro VV *změnit (se) k lepšímu* existují dvě různé TS.

Tato vlastnost odpovídá otázce „Nemůže metoda najít chybně něco navíc?“, její platnost pak vylučuje „false positives“.

Hypotéza C jistě nebude platit dokonale. Můžeme si vymyslet kolokaci,[28] která je lexikálně i syntakticky shodná s VV, a přesto ji za VV nelze považovat, neboť její význam je doslovný (zatímco VV, který se z doslovné kolokace zhusta vyvinul, má význam posunutý). Zjistit, do jaké míry tato vlastnost platí, je také cílem experimentu popsaného dále v sekci 6.1.6, k němuž si připravujeme podklady. Jak jsme ovšem uvedli již v úvodu, domníváme se, že takovéto přetěžování VV je

[28] Například *„náměstí působí nevyváženě kombinací* vysoké školy *a nízké radnice“*, *„ledva se vyšvihl na surf, smetla ho* nová vlna“, *„po čase jsem* došel k závěru, *povolil ho a pára unikla“*, nebo s větší představivostí můžu i *hodit flintu do žita*, aniž bych se vzdal.

v jazyce řídké, neboť (často záměrně) nabourává očekávání posluchače/čtenáře, což lze používat jen velmi střídmě.

Pro úplnost ještě dodejme, že doslovné užití není jediným důvodem pro „false positives“. Velice výjimečně může nastat případ, kde identifikace selže kvůli nedostatku informací uchovávaných v TS. Příkladem je věta *„Leonardo dal svým gólem signál k výhře.“* — [PDT, zkráceno], kde TS bez přidání například funktorů (zde *gól*$_{\mathsf{MEANS}}$) bude chybně identifikován VV *dát gól*$_{\mathsf{CPHR}}$.

Unikátnost TS (hypotéza D). Požadavek na unikátní TS ve slovníku úzce souvisí s hypotézou C, neboť porušení každé z nich vede k chybnému ohodnocení – v případě C k „false positives“, v tomto případě je sice VV nalezen ve správném rozsahu, ale je chybně určeno, o jaký VV se jedná.

Výjimek je nepatrné množství, zatím jsme objevili pouze VV *přímá volba*, která znamená jak výraz z oblasti volebního práva (*přímá volba prezidenta*), tak telekomunikační termín (*telefon s přímou volbou*) a *nová vlna*, což je buď umělecké hnutí, nebo druh ovčí vlny.

Vlastnosti B (jediná TS pro lexii) a A (souvislý podstrom) představují hlavní motivaci, proč se domníváme, že automatická analýza VV by měla na t-rovině fungovat lépe než na m-rovině – zatímco v povrchovém vyjádření se mohou blízko k sobě dostat slova, která spolu nesouvisí, a naopak jednotlivé části skutečného VV mohou být rozmístěny ve větě i velmi daleko od sebe a v různém pořadí, na t-rovině by se nic podobného stávat nemělo.

Ačkoli vlastnost C (žádné „false positives“) zdaleka neplatí stoprocentně, očekáváme, že mezi syntakticky shodně strukturovanými výrazy bude mylně označených VV méně než mezi všemi slovy věty analyzované na m-rovině.

Dále vlastnost A je zárukou, že je možné všechny t-uzly víceslovného výrazu ve stromové reprezentaci „sbalit“ a nahradit jediným uzlem pro celý VV, což bylo jedním z cílů projektu Lexemann (sekce 6.1.3) a dnes je to jeden ze zobrazovacích módů PDT 3.0.

6.1.6 Automatické vyhledávání tektogramatických struktur

Dostáváme se k samotné automatické identifikaci VV ze slovníku v textu. Popíšeme tři způsoby, jak toho dosáhnout, a tyto tři experimenty provedeme na třech datových sadách (datasetech).

Experimenty na třech rovinách jazykového popisu

Jednotlivé roviny PDT jsou mezi sebou provázané (viz sekce 3.3), takže se slovníkem SemLex spjatým s t-rovinou můžeme snadno pracovat i na zbylých rovinách. Srovnávací experimenty jsme tedy provedli na všech třech základních rovinách:

- Na t-rovině s využitím tektogramatických stromových struktur (t-TS) tak, jak jsou uloženy v SemLexu.
- Na a-rovině s využitím analytických TS (a-TS), které získáme z dat.
- Na m-rovině pouhým vyhledáváním lemmat s pevně danou šířkou okénka.

První experiment bude čistě syntaktický a bude probíhat na **t-rovině**. Všechny VV, které jsou v SemLexu, resp. jejich TS, budeme hledat v hloubkových stromech t-roviny.[29] V případech, kdy byla porušena vlastnost B a získali jsme z dat více různých TS, byla do slovníku uložena a následně použita nejčastější z nich. Očekáváme, že použitím t-roviny odstraníme problémy se slovosledem, s roztržením VV na víc částí ve větě a vložením dalších slov, s částečnou elipsou kvůli koordinaci atp. S ideální tektogramatickou rovinou, jak je navržena ve FGP, bychom měli získat na této rovině nejvyšší *„precision“* (česky přesnost), neboť bychom neměli označit nic, co není VV. Výsledky z t-roviny chceme srovnat s podobným přístupem na ostatních rovinách.

Druhý experiment na **a-rovině** je obdobou toho na t-rovině. Potřebné a-TS jsme získali z dat a-roviny PDT tak, že jsme do ní projektovali patřičné t-TS včetně všech pomocných slov. (Každý t-uzel obsahuje odkazy na odpovídající a-uzly: maximálně jeden pro autosémantické slovo a případně další pro pomocná slova.) Místo `t_lemma`t ukládáme obyčejná lemmata z m-roviny (`lemma`ta). Na a-rovině byla podle očekávání generalizace ještě nižší a jedna VV zde měla více a-TS. Všechny extrahované a-TS jsme uložili včetně jejich frekvencí do varianty slovníku (říkejme mu a-SemLex). Všechna data PDT pak procházíme stejným způsobem jako na t-rovině. Tentokrát na a-rovině hledáme přesnou shodu podstromu na a-rovině s a-TS a přesnou shodu `lemma`t. V případě většího počtu různých a-TS zatím využíváme (ve shodě s t-rovinou) pouze tu nejčastější.

Třetí experiment provedeme na **m-rovině**, kde však použijeme pouze pořadí slov ve větě a z veškeré morfologické informace pouze `lemma`. Zatímco experiment na t-rovině je z podstaty zaměřen na precision (nic, co není ve slovníku, se nenajde; ale co v něm je, mělo by se najít s vysokou spolehlivostí), zde si může-

[29] Technicky je vyhledávání paralelizováno přes všechna data PDT. V každé části jsou pak postupně vyhledávány všechny TS ze slovníku. Pro každou TS vybereme jeden uzel-list, který je ve struktuře nejníže položen, a kontrola zbytku TS nastává až ve chvíli, kdy je tento uzel nalezen při průchodu stromem. Postup k (obvykle jedinému) předchůdci ve stromě je rychlejší než prohledávání všech potomků, proto začínáme u nejnižšího listu. Kontroluje se přesná shoda `t_lemma`t a topologie podstromu. Prozatím nepoužíváme žádné další informace dostupné na t-rovině.

me nastavením velikosti „okna“ měnit poměr mezi precision a *„recall“* (úplností). Tím *oknem* máme na mysli počet sousedících slov v povrchovém zápisu věty, mezi kterými budeme VV hledat.[30] Experiment na m-rovině jsme tedy parametrizovali šířkou okna od dvou slov až po celou větu. Pracujeme pouze s lemmatizovanými tvary – kdykoli se všechna lemmata ze slovníkového hesla vyskytnou v okně, prohlásíme odpovídající slova za výskyt víceslovné lexie. V této verzi používáme nejjednodušší metodu, nehledíme tedy na pořadí slov a neměníme šířku okna v závislosti na délce VV, která může mít až dvanáct slov, viz tabulka 6.2. V nastavení s oknem širokým přes celou větu tato metoda pochopitelně „přegenerovává“ – nachází VV, které ve větě vůbec nejsou, jak jsme ukázali na příkladu v poznámce pod čarou 30.

počet slov	typy	instance
2	7063	18914
3	1260	2449
4	305	448
5	100	141
6	42	42
7	16	15
8	4	5
9	4	3
11	1	0
12	2	2

počet t-uzlů	typy	instance
1[31]	148	534
2	7444	19490
3	843	1407
4	162	244
5	34	32
6	13	8
7	3	1
8	4	1
9	1	1

Tabulka 6.2: Rozložení délky VV v SemLexu (typy) a v PDT 2.5 (instance). Vlevo měřeno počtem slov na povrchu, vpravo počtem uzlů na t-rovině.

[30] Příliš úzkým oknem šířky dva či tři nenajdeme *plnění plánu* v *„Plnění letošního odhlasovaného plánu se zbrzdilo.“*; s příliš širokým oknem šířky šest ho chybně najdeme i ve větě *„Nízké pojistné plnění vyplácené pojišťovnou nevyhovuje našemu plánu na obnovu.“*. Univerzální šířka okna neexistuje, což je vidět na větě *„Pojistné plnění našemu plánu rozhodně nevyhovuje.“*, kde i šířka tři vede k chybě.

[31] Jednouzlové VV existují a ačkoli nebylo naším cílem je anotovat, ne vždy bylo při anotaci zřejmé, že reprezentace VV na t-rovině bude mít jen jeden uzel, proto se do slovníku některé dostaly. Příkladem jednouzlového VV jsou výrazy s předložkou *na lavičce* a *bez váhání*; fráze *na správnou míru*, která je jako fráze zpracovaná už v PDT 2.0, s `t_lemma`tem „na_správnou_míru“; nebo slovesný výraz *umět si představit*, který je reprezentován jediným uzlem pro reflexivní sloveso „představit_si“ (a gramatémem deontické modality `deontmod` s hodnotou „fac“ místo modálního *umět*). Malá část těchto případů je také způ-

Testovací data

Navržené tři postupy jsme otestovali na trojích datech:

- na samotném PDT 2.0,
- na textech z PDT, které byly ovšem tentokrát anotovány automaticky, a
- na textech z ČNK, které byly automaticky anotovány stejným způsobem.

První dataset lze považovat nanejvýš za stanovení horní hranice úspěšnosti metody na datech této kvality, neboť používá stejná trénovací i testovací data. Z PDT 2.0 vznikl SemLex: měl by tedy pokrývat všechny VV obsažené v PDT a navíc uložené TS jsou z týchž ručně anotovaných dat t-roviny.

Druhý dataset vznikl zcela novou automatickou anotací textů PDT systémem Treex, jak byla popsána v sekci 6.1.4. Pomocí Treexu získáme m-rovinu, a-rovinu a t-rovinu, které potřebujeme pro tři experimenty. Tím simulujeme, jak by se naše metoda chovala na nových datech, která také nebudeme mít zpracovaná ručně. Velkou výhodou je, že pro PDT máme testovací data. Na druhou stranu tu zbývá problém se slovníkem, neboť SemLex je právě z těchto dat sestaven a neměl by v něm žádný VV chybět. Simulujeme tím tedy nereálnou situaci, kdy máme k dispozici kompletní slovník všech VV, žádná nová nás nepřekvapí. S touto výhradou jsou však výsledky poměrně vypovídající.[32]

Třetí dataset pochází z ČNK, z korpusu publicistiky SYN 2006 PUB. Pro syntaktickou anotaci na a-rovině i t-rovině jsme použili stejnou sérii bloků v Treexu jako u druhého datasetu. V těchto datech jsou konečně také výskyty nových výrazů, které při anotaci nebyly spatřeny a tudíž ve slovníku chybí. Třetí dataset představuje korektní testovací data, která se neshodují s textem, z něhož jsme tvořili slovník, a která jsou automaticky anotována stejným způsobem, jako to můžeme udělat s libovolným jiným textem. Nevýhodou je jen to, že pro tyto nové texty nemáme srovnatelně velká testovací data. Pro potřeby evaluace jsme ručně anotovali 546 vět.

Vyhodnocení

Tabulka 6.3 shrnuje výsledky všech tří experimentů na třech datasetech. První široký sloupec odpovídá původnímu, ručně anotovanému PDT 2.0, druhý automaticky zpracovaným datům PDT a třetí novým automaticky zpracovaným datům z ČNK. Sloupce obsahují precision (P), recall (R) a F_1 score (F),[33] vše

sobena aktuální elipsou, o které jsme psali v bodě 2 na straně 152, pokud v datech nemáme jiný než elidovaný výskyt.

[32] Je tu však ještě jeden problém: parsery, které jsme použily, jsou natrénovány právě na PDT, neboť jiná ručně anotovaná data k dispozici nejsou.

[33] F_1 score je definováno jako harmonický průměr precision a recall.

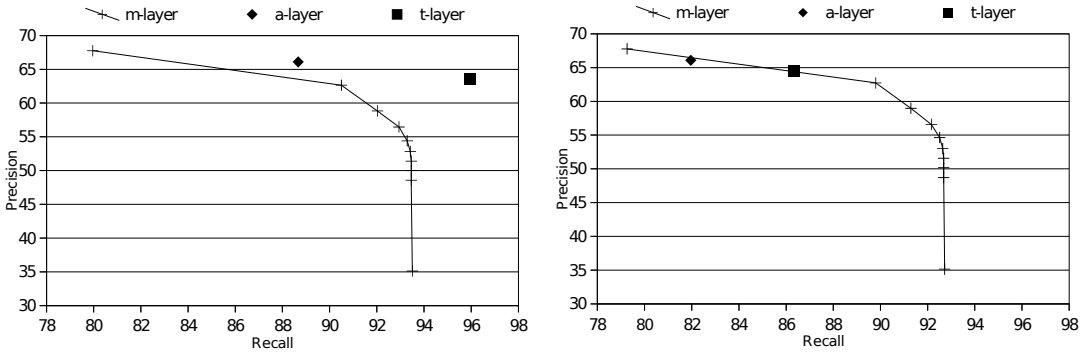
v procentech. První dva široké sloupce jsou výsledkem porovnání s „gold daty“ získanými z projektu Lexemann a výsledky jsou také vyneseny na obrázku 6.11. Třetí sloupec byl porovnán s 546 ručně anotovanými větami.

rovina	PDT (ručně)	PDT (automaticky)	ČNK (automaticky)
t	63,4 / **96,0** / **77,0**	64,8 / 86,3 / **74,7**	44,4 / 58,0 / 50,3
a	66,1 / 88,7 / 75,8	66,1 / 82,0 / 73,2	45,2 / 60,0 / 51,6
m / 2	**67,8** / 80,0 / 73,4	**67,8** / 79,3 / 73,1	**51,9** / 56,0 / **53,9**
/ 3	62,7 / 90,5 / 74,1	62,7 / 89,8 / 73,9	47,0 / 60,0 / 52,7
/ 4	58,8 / 92,0 / 71,8	59,0 / 91,3 / 71,7	42,8 / 61,3 / 50,4
/ 5	56,5 / 92,9 / 70,3	56,6 / 92,2 / 70,1	40,1 / 61,3 / 48,5
/ 6	54,5 / 93,3 / 68,8	54,6 / 92,5 / 68,7	38,3 / 61,3 / 47,1
/ 7	52,9 / 93,4 / 67,5	53,0 / 92,6 / 67,4	37,0 / 61,3 / 46,2
/ 8	51,4 / 93,5 / 66,3	51,6 / 92,7 / 66,3	35,6 / 61,3 / 45,0
/ 9	50,0 / 93,5 / 65,2	50,2 / 92,7 / 65,1	34,7 / 61,3 / 44,3
/ 10	48,6 / 93,5 / 63,9	48,7 / 92,7 / 63,9	33,8 / 61,3 / 43,6
/ ∞	35,1 / 93,5 / 51,1	35,2 / **92,7** / 51,0	22,7 / **62,0** / 33,2
	P / R / F	P / R / F	P / R / F

Tabulka 6.3: Vyhodnocení všech tří experimentů na třech sadách dat. Třetí sada byla kvůli nutnosti ruční evaluace výrazně menší než předchozí. Čísla jsou udávána jako procenta precision, recall a F_1 score. Deset variant experimentu na m-rovině se liší velikostí okna od dvou slov (2) po celou větu (∞). Nejvyšší číslo ve sloupci je vždy vyznačeno tučně (ačkoli v případě vysokého recall a mizivého precision nejde o použitelné výsledky).

Jednotlivé řádky tabulky 6.3 odpovídají třem experimentům: na t-rovině, na a-rovině a deset variant experimentu na m-rovině s různou šířkou okna od dvou do deseti a přes celou větu/souvětí.

Ruční anotaci vět z ČNK prováděl jediný anotátor s použitím stejného anotačního nástroje jako v původním projektu. Tento anotátor byl přítomen celému projektu Lexemann a znal anotační instrukce (byl jejich spolutvůrcem), nicméně nebyl jedním z původních anotátorů. V testovacích datech vyznačil 163 výskytů VV, z nichž 46 nebylo obsaženo v slovníku SemLex. To jsou tedy výskyty, které žádný z našich tří experimentů z podstaty nenalezne. (Škodí tedy všem stejně, což ale nevadí, neboť kvalita slovníku není předmětem našeho testování).



Obrázek 6.11: Výsledky z tabulky 6.3 vynesené do grafu pro precision a recall. Levý graf zobrazuje ručně anotovaná data PDT, pravý graf automatickou anotaci týchž dat.

Před vyhodnocením jsme museli přijmout dvě rozhodnutí: jak započítat jen částečný překryv s „gold daty“ a jak pracovat se zanořenými víceslovnými výrazy.

Když se vyskytne elipsa (bod 2 na straně 152) v trénovacích datech (a tudíž v SemLexu), nastává riziko, že se bude rozcházet výsledek automatické identifikace a „gold dat“.[34] Rozhodli jsme se nepenalizovat naši proceduru, pokud označí zkrácený výraz.

Při ruční anotaci PDT v projektu Lexemann jsme nepovolovali zanořené ani křížící se anotace – anotátoři si vždy museli vybrat. Při automatické identifikaci se však nijak nerozhodujeme, které VV zachovat. Vše, co je nalezeno, vyhodnocujeme. Tím nám sice vzroste recall, ale víc nám klesne precision.[35]

[34] Například řekněme, že máme ve slovníku pod heslem *ministerstvo průmyslu a obchodu* uloženu pouze elidovanou TS reprezentující *ministerstvo průmyslu*. Důvodem je, že se kompletní výraz v trénovacích datech neobjevil (či byl méně častý). Když potom na tuto kompletní verzi narazíme v testovacích datech, odlišně od kontrolní správné anotace vyznačíme jen podmnožinu, a přidělíme jí plný název ministerstva. Je toto plnohodnotná chyba? Je přiřazena správná položka, na dané místo je poukázáno – jen rozsah značky je odlišný.

[35] Jedna ze stálých otázek během ruční anotace se týkala správného zpracování názvů zákonů a s nimi spojených procesů. Když se pak v testovacích datech vyskytne například spojení *vládní návrh novely zákona o dani z přidané hodnoty*, můžeme při správně fungující anotaci vyznačit naráz všechny následující VV: *vládní návrh*, *novela zákona*, *zákon o dani z přidané hodnoty*, *daň z přidané hodnoty*, *přidaná hodnota* a navíc ještě celý výraz dohromady, tedy celkem šest možností. Zatímco se jedním z nich zřejmě trefíme do anotace v „gold datech“, bude to za vysokou cenu: tento samotný úsek textu by měl recall 100 %, ale precision jen 17 %.

6.1.7 Diskuse výsledků a výhled do budoucna

Je zajímavé, že výsledky v prvním a druhém sloupci v tabulce 6.3 se příliš neliší. Téměř shodná čísla na m-rovině však nejsou velkým překvapením, neboť automatické morfologické značkování (95,68 % – Spoustová, 2008) dosahuje úspěšnosti ručního. Pracujeme tedy na m-rovině s téměř stejnými lemmaty v obou datasetech.

Na vyšších rovinách je důvod méně zřejmý. Největších rozdílů (nízkých 10 procentních bodů) dosahuje recall, zbytek se příliš neliší. Domníváme se, že důvodem může být druh informací, které od parserů vyžadujeme. Automatické parsery zejména pro t-rovinu řeší komplexní sadu anotačních úkolů (od struktury stromu včetně odebrání či přidání uzlů přes pořadí až po desítky informací pro každý uzel) a jejich úspěšnost proto stále není ideální. Nicméně malá část z těchto informací (struktura a `t_lemma`ta), kterou potřebujeme my, je zřejmě z těch snazších, a proto na relativně dobré úrovni. Složitější věci, které také využíváme, jsou relativně řídké (koordinace), případně jsou nedostatečné a s chybami i v ruční anotaci (doplnění uzlů za elidovaná slova) či dosud nejsou implementované (jednotná `t_lemma`ta – viz hypotéza B). Dalším důvodem bude již zmíněné natrénování parseru na stejných datech.

Výsledky (tabulka 6.3 a obrázek 6.11) nepotvrdily, že v současném stavu je nejvýhodnější identifikaci provádět na t-rovině, mj. i proto, že je ještě poměrně vzdálená ideální rovině jazykového významu. F_1 score pro jednotlivé roviny se ovšem nikdy neliší o mnoho a výsledky na vyšších rovinách jsou často proti m-rovině lepší alespoň pro precision či recall. Nezdá se tedy, že m-rovina by byla plnou odpovědí, neboť nějaká syntaktická informace zřejmě bude potřebná.

Rádi bychom v budoucnu snížili vliv obou příčin, které zhoršují výsledky automatické identifikace na t-rovině, tedy použitý slovník (a metoda) a vlastnosti t-roviny samotné.

Vylepšení t-roviny

Začneme současným stavem t-roviny. Ukázalo se, že t-rovina není dostatečně abstraktní k tomu, abychom se mohli spolehnout, že jakožto rovina jazykového významu odstíní většinu „nevýznamových“ drobností a vyřeší za nás záležitosti spadající na nižší roviny jazykového popisu. V podstatě se t-rovina od a-roviny liší (pro dané jevy) mnohem méně, než bychom čekali a potřebovali, čemuž odpovídají i výsledky. Většinu našich výhrad a možných nedostatků t-roviny jsme vyjmenovali na předchozích stránkách, v části Vlastnosti reprezentace VV pomocí TS. V některých ohledech by pomohlo, kdyby se t-rovina přiblížila svému předob-

razu, abstraktnější tektogramatické rovině FGP, v jiných ohledech navrhujeme řešení, která vedou ke sjednocení `t_lemma`t nad rámec FGP.

Dále do otázky nedostatků t-roviny vstupuje automatická syntaktická analýza. Ačkoli jsme řekli, že většinu informací, které dokážeme využít, zřejmě poskytuje v rozumné kvalitě, ještě jí mnoho schopností chybí, například nedoplňuje správně t-uzly pro elidované části koordinací apod. Zřejmě má také horší výsledky na neznámých datech ČNK.

Rádi bychom provedli některé z lexikálních změn navržených na stranách 152–155 alespoň jako „postprocessing“ již hotových (ručních či automatických) syntaktických anotací a ověřili, zda a jak pomůže takové sjednocení `t_lemma`t automatické identifikaci. Například pro vidové dvojice by mělo být možné sjednocení u většiny sloves provést automaticky,[36] zatímco třeba zdrobněliny, přechylování či slovní druhy by vyžadovaly buď ruční práci, nebo využití derivačního slovníku, například slovníku DeriNet (Ševčíková a Žabokrtský, 2014).

Takovéto zpracování `t_lemma`t – jejich zobecnění – by nebylo užitečné výhradně pro víceslovné výrazy. Domníváme se, že by mohlo pomoci i dalším úlohám zpracování přirozeného jazyka. Již nyní se například při překladu pomocí systému Treex používají derivační slovníky, v nich ovšem přechylování, vidové dvojice, ani zdrobněliny zahrnuty nejsou. Přechýlení ani zdrobňování nejsou sice jevy natolik časté, aby jejich vyřešení mělo zásadní dopad na kvalitu překladu, částečně pomoci by to ale mohlo. Bylo by pak dobré experimentálně ověřit, zda schopnost přejít z neznámého („out-of-vocabulary“) slova ke slovu derivačně příbuznému a známému (spolu se znalostí druhu té derivace) je ku prospěchu věci. Riziko naopak je, že spojení více slov pod jediné `t_lemma` může překladový model zamlžit.

↪ *Příklad:* Řekněme, že systém nezná slovo *kytoveček*. Pokud by ale mohl (s využitím DeriNetu, nebo pravidlové heuristiky) slovo převést na `t_lemma` $kytovec_{\texttt{dimin}}$, které už do angličtiny přeložit umí, zbývá otázka, zda se dokáže naučit, že příznak „dimin“ se překládá předřazením nového slova *little*.

Vylepšení SemLexu

Druhou oblastí, kterou určitě budeme vylepšovat díky zkušenostem z těchto experimentů, je slovník víceslovných výrazů a informace v něm uložené a způsob jejich využití při identifikaci na t-rovině. Začínali jsme s představou, že takto jednoduchá TS – obsahující pouze závislostní strukturu a `t_lemma`ta autosémantických uzlů – bude stačit. Je zřejmé, že je nutné slovník obohatit a v souvislosti s tím mírně změnit i vyhledávací algoritmus.

[36] Slovesa již mají svůj gramatém vidu vyplněný, bylo by tedy pouze potřeba nalézt a sjednotit odpovídající si protějšky. K tomu můžeme využít VALLEX, který právě tuto informaci pro nejčastější česká slovesa poskytuje.

- Přinejmenším bude potřeba zahrnout **synsémantická slova**, která jsou součástí VV. K tomu bude potřeba umět je odlišit od těch, která se jen vyskytnou v okolí VV.
  ↪ V současnosti nedokážeme odlišit *na svou pěst*, *do svých pěstí* a *svou pěstí* (tedy i část druhého a třetí výraz budeme mylně považovat za VV) – ze stejného důvodu jako správně nerozlišujeme *na přenosovou soustavu*, *do přenosových soustav* a *přenosovou soustavou*, kteréžto výrazy všechny obsahují (tentýž) VV.
  Budeme tedy muset rozlišit například tyto předložky uvozující VV jako větný člen od předložek, které jsou jeho součástí, a podle toho upravit položky ve slovníku a při identifikaci se podle toho také chovat. Kromě drahé ruční úpravy máme dvě možnosti: využít porovnání BASIC_FORM (která pomocná slova obsahuje) s `t_lemma`ty v TS, nebo z dat odhadnout pomocná slova, která se nemění a vyskytují se ve všech instancích, které v PDT máme.
- Podobným způsobem (kontrolou počtu synsémantik v BASIC_FORM a v TREE_STRUCTURE) bychom mohli vytipovat všechna slovníková hesla, která jsou **neúplná** kvůli elipse v trénovacích datech, a TS buď rozšířit, nebo doplnit o další varianty (viz níže).
- V mnohem menší míře jsou zastoupené případy, kde jde o v nějakém směru **ustrnulý VV**, přesto by bylo dobré v budoucnu umět i toto v SemLexu zachytit (může jít o omezení pouze na plurál, výhradně vybrané pády, případně dokonce o „šablonovité“ VV, které umožňují jen přesně definované obměny,[37] apod.)
  ↪ Vždy je jen *čistírna odpadních vod*, nikdy **čistírna odpadní vody*. Výrazu *ptačí perspektiva* lze užít bez předložky, ale v předložkovém užití není zřejmě jiná možnost než předložka *z*.[38] Naprosto žádná variace není povolena ve výrazu *na plné obrátky*: ten je zcela zafixován a používá se pouze jako pevný řetězec. Vždy pouze *neomezené možnosti*, bez negace to použít nelze (resp. pak to již není VV).

[37] Máme na mysli výrazy jako *„*** si na dno sil“*, kde na místě hvězdiček mohou vystupovat pouze dvě slovesa: *sáhnout*, nebo *hrábnout*. Kdežto před takovým *volnou ruku* už může stát mnohem víc sloves, stále však zdaleka ne všechna: *dostat/mít/potřebovat/požadovat / dát/poskytnout/nechat/ponechat volnou ruku*, případně možná ještě *slíbit*, pokud to není spíše elipsa ze *slíbit dát volnou ruku*.

[38] Ověřováno na všech 146 výskytech kolokace *ptačí perspektiva* v SYN 2010. Naprostá většina užití byla s předložkou, a to s předložkou *z*, výjimku tvořily dva výskyty předložky *s*: jednou jako knižnější varianta předložky *z* ve stejném významu, podruhé v elipse, kterou hodnotíme až jako chybu: *„Nechci vás nudit popisováním panoramatu známého z tisíců pěkných rytin, s ptačí perspektivou města Gdaňsku.“* — [Plechový bubínek, překlad Vladimír Kafka, Josef Hiršal, vydáno 2001]

- Již v souvislosti s hypotézou C jsme narazili na potřebu (ačkoli poměrně výjimečnou) využívat informaci o **dalších atributech** t-uzlů, jako je `functor` (*dal gól* vs. *dal gólem signál*), nebo o technických uzlech jako je RHEM #NEG (zmíněné *neomezené možnosti*).
- Předposlední úprava slovníku už nesouvisí se strukturou TS, ale s jejich množstvím. Vysvětlili jsme důvody, proč může být porušena hypotéza B, tedy jak se stane, že jedna víceslovná lexie může mít více podob v t-stromě, a tedy **více TS**. Prozatím jsme to vyřešili tak, že jsme ostatní struktury ignorovali a do slovníku uložili jen jedinou, nejčastější. Hypotéza B byla ovšem porušena u 8,75 % slovníkových hesel – víme tedy, že každé z nich má v datech nejméně jednu instanci, která vypadá jinak než naše TS. Bylo by potřeba vyzkoušet, je-li to řešení nejlepší – zda vložení všech ostatních, často poškozených TS (různě elidovaných či jinak obměněných) a vyhledání všech takových nepřinese dostatečné zvýšení recall. Pochopitelně nejen na t-rovině, stejný přístup můžeme aplikovat i na a-rovinu.
- Ruku v ruce s tím by muselo jít řešení **překrývajících se VV**. Pokud si do slovníku pro jeden VV uložíme všechny TS, tedy i elidované, označíme často stejný výskyt s více různými překrývajícími se rozsahy. Z nich bude samozřejmě správný maximálně jeden a zbytečně si snížíme precision. Tuto úpravu by ovšem bylo užitečné vyzkoušet i bez povolení více TS.

Navržené změny by měly vylepšit výsledky na t-rovině (případně i na a-rovině) ve srovnání s m-rovinou.

Není ale nutné držet se za každou cenu t-roviny. Vzhledem k překvapivě dobrým výsledkům na a-rovině se nabízí možnost použít na datech, která jsou beztak mezi rovinami provázaná, postup neomezující se na jedinou rovinu a používat informace z obou vyšších rovin. To jsme v podstatě již navrhli v předchozích odstavcích, kde doporučujeme kombinovat t-rovinu s pomocnými slovy z a-roviny, na která vedou z t-uzlů odkazy. Protože kvalita automatické syntaktické analýzy na a-rovině je vyšší, měli bychom vyzkoušet, zda i jiné její vlastnosti nemohou být pro nás užitečné.

Pro úplnost dodejme, že zatímco vyšší roviny těží z bohaté struktury anotace, ale vyhledávací metoda je relativně jednoduchá, a proto i vylepšování se týká zejména trénovacích dat a slovníku – na jednoduše anotované m-rovině jsou možnosti vylepšování vyhledávací metody téměř neomezené. Nezmiňujeme se o nich, neboť nejsou naším cílem, m-rovina nám sloužila pouze pro srovnání.

Závěr

Identifikace konkrétních VV v textu dosud nebyla příliš široce prozkoumána. Nicméně její důležitost pro další úlohy zpracování přirozeného jazyka je zjevná.

Všude, kde potřebujeme pracovat i s významem nalezených VV (např. vyhledávání informací, sumarizace textu, překlad) se dodatečné informace ve slovníku mohou hodit. Náš slovník SemLex k tomu skýtá ideální zdroj, neboť obsahuje téměř 9 000 VV, všechny se syntaktickou strukturou extrahovanou z dat a s dalšími informacemi.

VV z tohoto slovníku jsme se pokusili identifikovat postupně v trojích datech: v původním, ručně anotovaném PDT, v datech PDT anotovaných automaticky a stejně tak automaticky anotovaných nových datech ČNK.

Provedli jsme experiment na t-rovině a pro srovnání také na a-rovině a m-rovině. Převaha přístupu skrze t-rovinu se neprokázala, mj. z důvodů nedostatečné reprezentace VV ve slovníku, ne dosti rozsáhlé generalizace `t_lemma`t v PDT a určitých vad automatické syntaktické anotace.

6.2 PCEDT a entity BBN

V předchozí sekci 6.1 ČNK a SemLex jsme popsali metodu vyhledávání VV s využitím hloubkové syntaktické anotace. Konkrétně jsme hledali **české** víceslovné **lexie** z textů PDT. Zde vyzkoušíme stejný přístup v pozměněných podmínkách, abychom otestovali konzistenci anotace a zároveň ověřili širší užití metody. Budeme vyhledávat **anglické** víceslovné **pojmenované entity** z textů Wall Street Journalu.

Cílem je zejména kontrola konzistence anotací – zda stejné entity vypadají vždy stejně a jsou vždy správně anotované. Neklademe si za cíl identifikovat všechny entity, chceme se naopak věnovat (zdánlivým) chybám a ověřit, na které úrovni chyba nastala (Penn Treebank, BBN, PCEDT – viz dále – či náš algoritmus).

Dále chceme vyhledávací algoritmus (popsaný v předešlé sekci, v 6.1.6) vyzkoušet v poněkud odlišném prostředí, ověřit užitečnost využívání hloubkově syntaktické informace pro tento druh úloh a získat další podněty pro jeho úpravy.

Nejprve představíme data, se kterými budeme pracovat (6.2.1), následně popíšeme úpravy algoritmu (6.2.2) a vyhodnotíme výsledky – ve smyslu úspěšnosti algoritmu, ale zejména ukázkami nalezených nekonzistencí anotace, které budeme chtít opravit (6.2.3). Sekci uzavřeme výhledem do budoucna (6.2.4).

6.2.1 Gigantum super humeris

Nejprve stručně představíme jazykové zdroje, které jsme využili. Vše, co jsme pro tento experiment potřebovali, už bylo připraveno, stačilo to jen použít – postavit

se „na ramena obrů“. Jedná se o PCEDT (potažmo Penn Treebank) a anotaci entit tamtéž od společnosti BBN.

Od WSJ k PCEDT

K dispozici je zdigitalizovaný anglický novinový text Wall Street Journalu (WSJ), který je součástí Penn Treebanku (sekce 3.2). Složková syntaktická anotace Penn Treebanku byla doplněna o vnitřní strukturu jmenných frází (Vadas a Curran, 2007). Na základě desetiletí rozvíjené teorie FGP byla tato anotace převedena do hloubkové závislostní syntaxe: vznikla pro ni t-rovina a celý korpus PCEDT (sekce 3.5). Tím jsme získali pro nás potřebnou anotaci anglického textu ve stejném tvaru, v jakém jsme pracovali s českým textem v předchozí sekci.

Entity od BBN

Podle našeho zjištění neexistuje anotace víceslovných výrazů v datech WSJ. Je zde však k dispozici anotace pojmenovaných entit (a některých jmenných výrazů), z nichž mnohé jsou z podstaty víceslovné. Zaměříme se tedy pouze na ně.

BBN je americká technologická firma, která se zaměřuje obecně na výzkum a vývoj. V roce 2005 vydala anotaci entit v Penn Treebanku (v části s WSJ) s názvem BBN Entity Type Corpus.[39] Anotace pomocí vlastních nástrojů byla prováděna ručně a bylo označováno

- dvanáct typů pojmenovaných entit (PERSON, FACILITY, ORGANIZATION, GPE[40], LOCATION, NATIONALITY, PRODUCT, EVENT, WORK OF ART, LAW, LANGUAGE a CONTACT-INFO)
  ↪ *Příklad: James A. Talcott, Indiana Roof ballroom, Consolidated Gold Fields PLC, West Groton, River Danube, Southeast Asian, Scholastic Aptitude Test, World War II, Johnny B. Goode, 1988 trade act, Japanese language, 153 East 53rd St.*;
- deset typů jmenných označení, z nichž mnohé víceslovné bychom asi v projektu Lexemann nazvali víceslovnými lexiemi (PERSON, FACILITY, ORGANIZATION, GPE, PRODUCT, PLANT, ANIMAL, SUBSTANCE, DISEASE a GAME)
  ↪ *Příklad: vice president, grain elevators, savings institutions, nations, machine guns, Plantago ovata, black widow, gallium arsenide, Down's syndrome, lawn bowling*; a

[39] `https://catalog.ldc.upenn.edu/LDC2005T33`, vydáno společně s korpusem zájmenné koreference.

[40] geo-political entity

- sedm číselných typů, které v mnoha přístupech do kategorie pojmenovaných entit spadají (DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL a CARDINAL)
  ↪ *Příklad: 61 years old, the next morning, 8.47 %, $ 352.7 billion, about 321,000 barrels, more than a third, more than three.*

Mnoho z uvedených 29 typů se ještě dělí na podtypy (například FACILITY se dělí na FAC:AIRPORT, FAC:ATTRACTION, FAC:BRIDGE, FAC:BUILDING, FAC:HIGHWAY_STREET, FAC:HOTEL, FAC:OTHER); celkem se v anotacích nachází 105 podtypů. Ty byly tedy použity pro anotaci WSJ a v celém korpusu se nachází 171 880 výskytů BBN entit.

Dále jsme měli k disposici upravenou verzi PCEDT od Josefa Tomana, na jejíž t-rovinu byly entity od BBN vloženy. Z nich jsme vybrali víceslovné (přesněji ty, které obsahovaly více než jeden t-uzel) a pouze s těmi pracovali. (Proto jsme i výše uváděli pokud možno víceuzlové příklady.) Celkem jsme extrahovali 19 036 entit,[41] které se ve WSJ vyskytovaly v celkem 57 268 případech. Kategorizovány byly do 93 podtypů (z oněch 105 podtypů používaných BBN).

6.2.2 Identifikace BBN entit

V této sekci popíšeme metodu, kterou jsme použili pro identifikaci entit v PCEDT. Použili jsme PCEDT se zanesenými anotacemi BBN entit od Josefa Tomana, které jsme použili jako trénovací data. Tyto entity jsme se pokoušeli znovu vyhledat v datech, odkud jsme je odstranili. Metoda je obdobou metody použité pro české víceslovné lexie (sekce 6.1.6), proto ji popíšeme zejména z hlediska změn, které ji odlišují.

Lexikon BBN entit

Ze všeho nejdříve jsme si museli vyrobit „slovník použitých entit“. Prošli jsme veškerá data PCEDT a entity, které obsahovaly více uzlů, jsme si uložili do slovníku ve stejném formátu jako má SemLex (sekce 4.9). Zejména tedy jejich TREE_STRUCTURE (TS) a TYPE BBN entity, ale také jsme např. extrahovali BASIC_FORM, přidělili ID, či spočítali WSJ_FREQ. Vzhledem ke způsobu extrakce entit z PCEDT (viz dále sekce 6.2.4) jsou všechny nalezené a uložené entity z definice souvislé podstromy.

BBN entity, zejména číselné, se ovšem od lexií, s nimiž jsme pracovali dosud, výrazně liší v tom, že zde jsou rozsáhlé skupiny, které dodržují stejný vzorec, přitom se však jedná o odlišné entity.

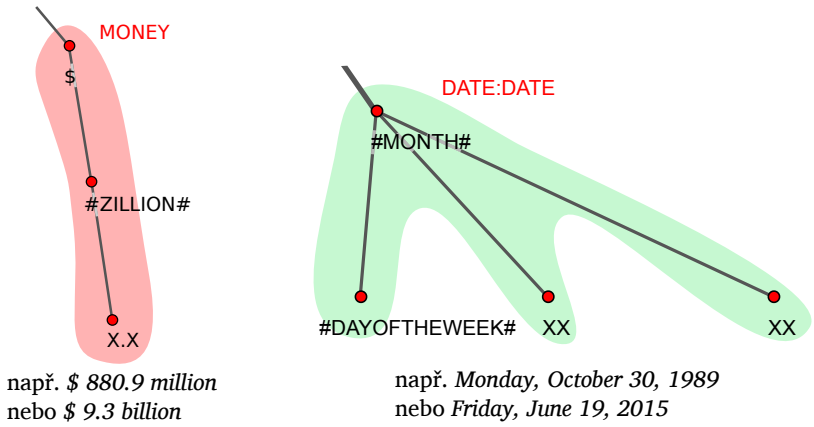
[41] Vzhledem k sdružování více entit pod jednu šablonu (viz sekce 6.2.2) se pod tím číslem ve skutečnosti skrývá 28 520 odlišných entit.

↪ *Příklad: 1.85 cents; 11.5 cents; 13.79 cents; …*
Nov. 9 1994; March 31, 1994; MAY 1, 1975; …
at least $ 110 million, at least $ 15 million, at least $ 7 billion

Provedli jsme tedy (pomocí naší vlastní funkce `unify_word()`) už při ukládání do slovníku následující substituce vybraných `t_lemmat` uvnitř větších entit:

- přirozená čísla → XX (*11*)
- dekády → XXXXs (*1960s*)
- desetinná čísla → X.X (*3.14*)
- tisíce → X,X (*12,500*)
- číslovky → #NUMBER# (*seven; eleven; twenty*)
- řády → #ZILLION# (*trillion*)
- měsíce → #MONTH# (*January; Feb.*)
- dny v týdnu → #DAYOFTHEWEEK# (*Saturday*)

Jedna položka slovníku pod sebou tedy může sdružovat stovky víceslovných entit (nejčastější typ je „$ <desetinné_číslo> <označení_řádu>“, tedy např. *$ 9.671 trillion*, jehož šablonovitou TS vidíme na obrázku 6.12 vlevo.)



Obrázek 6.12: Ukázky dvou „šablonovitých“ TS ve slovníku BBN entit. Levá TS je ve WSJ nejčastější, reprezentuje 1 613 výskytů. Pravá TS může také zastupovat stovky či tisíce dat, nicméně ve WSJ se vyskytuje jen jedenkrát.

Identifikace podle lexikonu BBN entit

Samotné vyhledávání je potom (mírně optimalizovanou) obdobou vyhledávaní TS podle SemLexu ze sekce 6.1.6, konkrétně popsané v poznámce pod čarou číslo 29 na straně 159.

Protože budeme opět existenci TS v t-stromě PCEDT ověřovat odspodu, zaindexovali jsme nejprve všechny položky lexikonu podle jejich nejnižšího uzlu. Při průchodu t-stromem nejprve pro každé `t_lemma` provedeme stejné substituce (`unify_word()`) jako při výrobě lexikonu (takže např. místo původních `t_lemma`t *35,600* a *11,455* spolu porovnáváme *X,X* a *X,X*). V dalším ověřování přítomnosti nějaké TS pokračujeme pouze tehdy, pokud se takto upravené `t_lemma` nachází v seznamu nejnižších uzlů.

Dále jsme zjistili z průběžných výsledků, že jsou tu (vedle už uplatněných substitucí) ještě jiné pravidelné varianty entit. Například pro tisíce se vyskytují tyto varianty (vždy jsou jako entita vyznačeny celé): *almost 10,000; approximately 565,000; at least 6,000; just over 5,000; just 10,000; more than 2,000; nearly 3,000; only about 10,000*. Z korpusu CzEng[42] (Bojar et al., 2012)[43] jsme pomocí dotazovacího jazyku PML-TQ získali anglická adjektiva a adverbia, která rozvíjí čísla. Z nich jsme sestavili dvě skupiny (upřesňujících, či rozvolňujících) „univerzálních atributů“: číselné atributy (*about, almost, approximately, around, equal (to), just, (at) least, less (than), more (than), nearly, next (to)*) mohou rozvíjet CARDINAL, PERCENT, MONEY, DATE:DATE a DATE:DURATION a časové atributy (*early, late, next, past, previous*), které rozvíjejí DATE:DATE a DATE:DURATION. Pokud se při identifikaci některý z vybraných univerzálních atributů vyskytne v okolí již nalezené TS (jako závislý uzel kdekoli, nebo jako řídící uzel celé TS),[44] připojíme ho do téže entity, ačkoli takto modifikovanou TS zrovna ve slovníku nemáme.

Kromě těchto slov je občas součástí entity také vyjádření postoje (*only, record, …*), které považujeme za chybu anotace, neboť na rozdíl od předchozích modifikátorů částku nijak nemění. Pro takováto slova tedy žádný seznam nepoužíváme.

V přidávání podobných pravidel by bylo možné pokračovat dále:

- Jakékoli číslo (CARDINAL) závislé na uzlu s `t_lemma`tem #PERCNT rozšíříme o tento uzel a změníme typ entity na PERCENT.
- Celý podstrom pod jedním ze slov *day, week, month, year, century, millenium* označíme jako DATE:DATE.
- S využitím slovníku nejčastějších křestních jmen můžeme odhadovat, že dvě po sobě jdoucí slova začínající velkým písmenem (jedno z nich ze slovníku) jsou PERSON.

[42] Z jeho zhruba jednoho procenta, které činí více než 150 tisíc vět, které je zveřejněno na `https://lindat.mff.cuni.cz/services/pmltq/czeng`.

[43] Protože chceme metodu evaluovat cross-validací, nemůžeme tato rozvití získat z některé části PCEDT. Tímto postupem samozřejmě některá rozvití do seznamu nevložíme, například *well over* a jiné.

[44] Jak ukážeme v sekci 6.2.3, takovýto univerzální atribut může být ve stromě umístěn celkem libovolně.

- Celý podstrom pod kořenem *Inc., Co.*, nebo *Corp.* označíme jako ORGANIZATION:CORPORATION. V omezené ekonomické doméně WSJ můžeme totéž odhadovat i pro podstromy s kořenem *&* (např. *Valley Federal Savings & Loan, Connecticut Bank & Trust*).
- V případě nalezení číselného typu zkontrolujeme, zda se nejedná o rozsah.
  ↪ Například nemusí jít o *175 million Canadian dollars*, nýbrž o *175 million to 180 million Canadian dollars*, což je složitá koordinační struktura na obrázku 6.13, kterou známe pro *16 to 19 years old* nebo *180 to 270 days*, ale řekněme, že s kanadskými dolary jsme ji v trénovacích datech neviděli.
- Podobný případ je rozsah, například *between 9% and 11%*.

Naším cílem však není ručně budovat pravidlový systém. Chceme naopak systém, který si umí informace získat z t-roviny samostatně, jen jsme mu v několika jednotlivých případech pomohli doplňkovým pravidlem.

Poslední věc, která se liší od práce s PDT, je zpracování podmnožin. Už z předchozích odstavců je zřejmé, že se snadno stane, že se ve větě *„The company produces almost 35 million barrels of oil a day.“* nalezne současně *35 million*, *almost 35 million*, *35 million barrels* a případně i celá anotovaná entita *almost 35 million barrels*. V takovém případě při dodatečném zpracování výsledků všechny podmnožiny (tedy první tři VV) jiného nalezeného VV odstraníme.

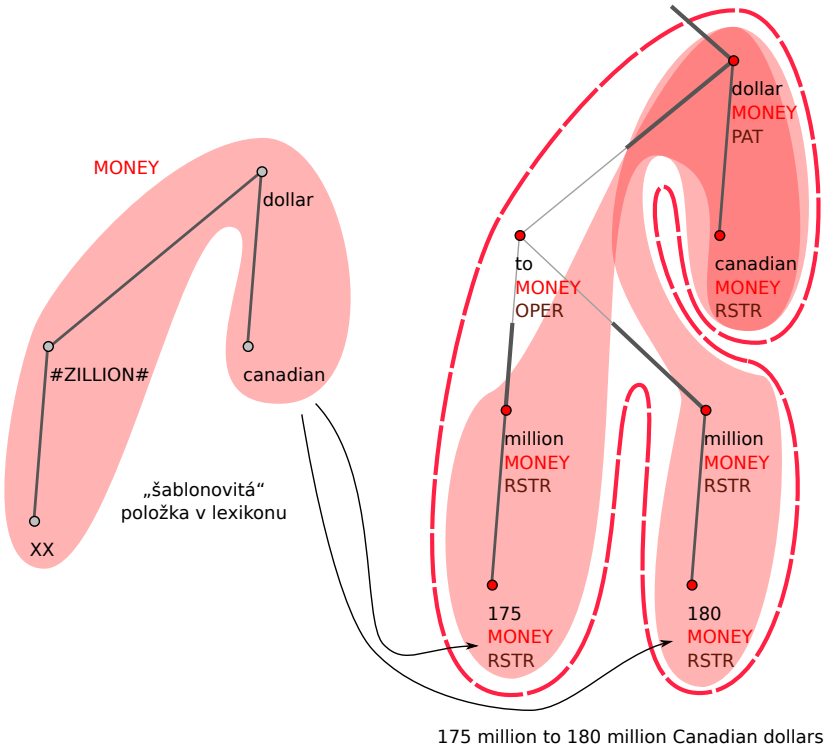
6.2.3 Vyhodnocení a nekonzistence anotací

Popisovaná metoda ze své podstaty nemůže nalézt VV, který se nevyskytl v trénovacích datech – a s výjimkou dvou drobných pravidel popsaných v sekci 6.2.2 (tedy substituce a univerzální atributy) se ani o nic podobného nesnaží. Předpokládáme, že výsledky budou obsahovat mnoho výrazů nenalezených ve slovníku (tzv. „out-of-vocabulary“ výrazů). Zajímají nás však zbylé chyby, které mohou ukazovat na chyby v anotaci BBN entit, či na chyby při převodu Penn Treebanku do t-roviny PCEDT. Evaluace tu slouží pouze k určitému zhodnocení samotné metody.

Použili jsme desetinásobnou cross-validaci, kdy jsme z PCEDT určili desetinu dat jako testovací a ze zbytku jsme sestavili lexikon. Ten jsme potom způsobem popsaným v předešlé sekci aplikovali na testovací desetinu a spočítali úspěšnost. Postup se desetkrát opakuje a testovacími daty se stávají jednotlivé desetiny, deset výsledků se zprůměruje.

Průměrná precision činí 69 %, recall 66 %. F_1 score je tudíž 67 %.

Mezi chyby počítáme i správně vyznačenou entitu s chybně přiřazeným TYPE a vyznačení podmnožiny entity (viz obrázky 6.13 a 6.14).



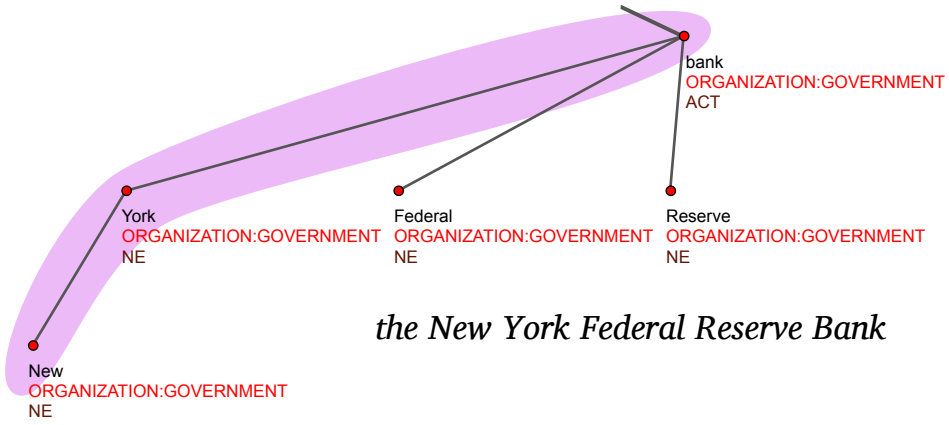
Obrázek 6.13: Identifikace entity MONEY uvnitř větší entity rozsahu. Vlevo vidíme slovníkovou TS pro například *39 billion Canadian dollars* a vpravo potom skutečný výskyt v datech, kde byla tato položka nalezena dvakrát. (Relací efektivního rodiče byla vynechána koordinační spojka *to*.) To je naznačeno částečně transparentní oblastí. Skutečná anotace od BBN (naznačena přerušovanou čarou) však zahrnuje celý rozsah, tedy obě částky i spojku.

Zároveň jsme podle očekávání narazili na nekonzistentní anotace, jejichž některé ukázky přinášíme. Díky této metodě je můžeme cíleně vytipovat a umožnit jejich opravu v příští verzi PCEDT.

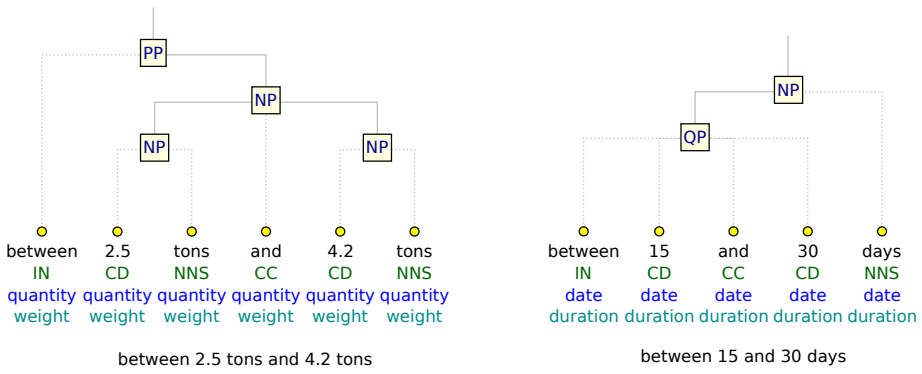
Ukázka nekonzistentní anotace Penn Treebanku:

- Vazba *between *** and ****. V Penn Treebanku může mít tato vazba dvě podoby: buď jako QP, tedy „Quantifier Phrase (i.e. complex measure/amount phrase)“, nebo je zpracována běžným způsobem. Není dodržováno ovšem zcela konzistentně, například pro dva kvantitativní rozsahy „tun“ a „dní“ je to vidět na obrázku 6.15.

Ukázka nekonzistentní anotace BBN entit:



Obrázek 6.14: Ukázka chybné identifikace – nalezena pouze podmnožina. Namísto plného názvu byla nalezena pouze *New York Bank*.

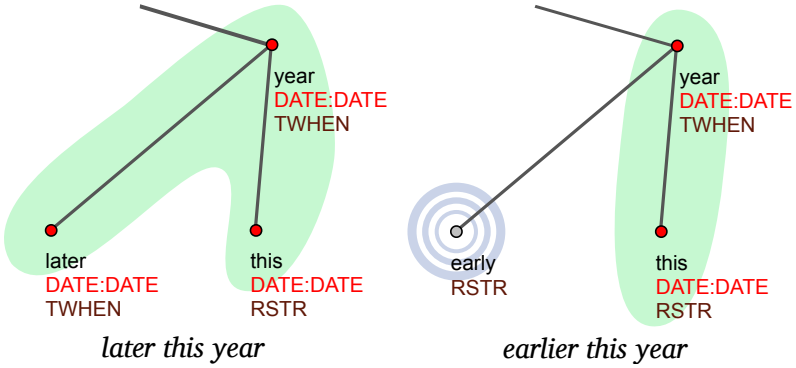


Obrázek 6.15: Dva obdobné rozsahy a jejich rozdílné zachycení v Penn Treebanku.

- Již zmíněné „univerzální atributy“ nejsou v anotacích BBN zahrnuty pokaždé, což považujeme za chybu, kterou by bylo dobré opravit, což s naším algoritmem nepředstavuje moc práce. Ukázka na obrázku 6.16.

Ukázka nekonzistencí na t-rovině PCEDT:

- Některá `t_lemma`ta čísel vyšších řádů jsou normalizována do tvaru s podtržítkem oddělujícím řády (*50_000*), jiná jsou ponechána v původním tvaru (*50,000*).
- „Univerzální atribut“ je, jak už jsme naznačili v předchozí sekci, anotován různě, což dobře dokládají čtyři varianty na obrázku 6.17. V uvede-

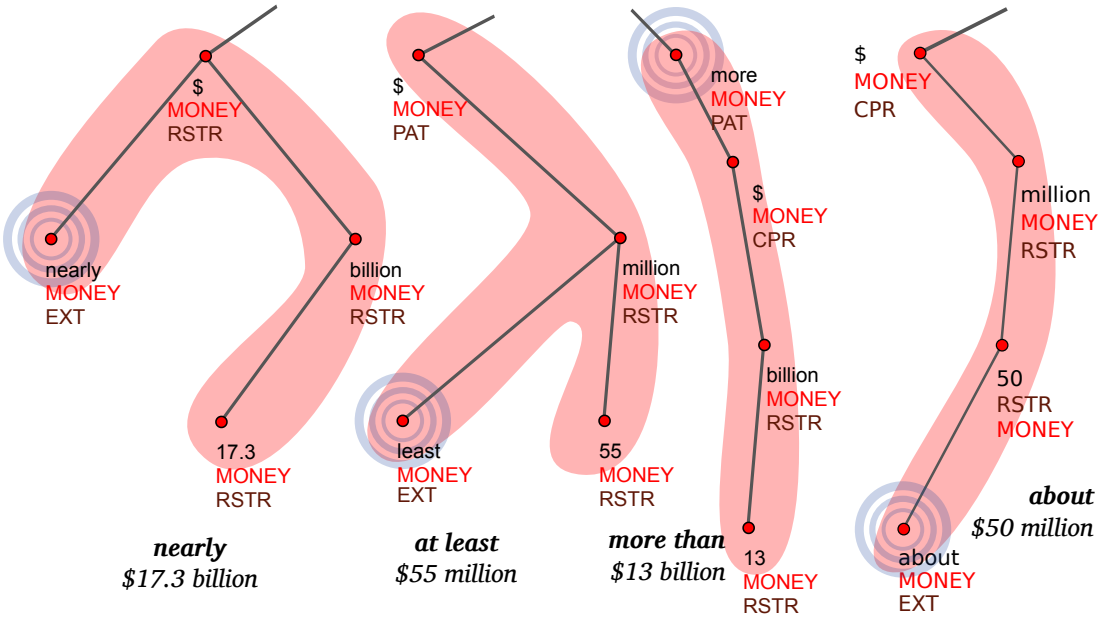


Obrázek 6.16: Výrazy *later/earlier this year* s odlišnou anotací BBN entit. V druhém případě není slovo *earlier* (zvýrazněno) součástí entity.

ných případech (*$50 million*) se domníváme, že vhodné místo pro tento atribut je rozvití číselného výrazu[45] (jako je tomu u druhého případu *at least* na obrázku 6.17), neboť jde o nějakou částku, k té se dodává upřesnění/„znepřesnění“ a takto upravená částka je pak v dolarech. (Totéž pro *almost 35%*, nebo *around 200,000 tons*.) I kdybychom však argumentovali pro zavěšení o úroveň výš, tedy jako rozvití až celé částky v dolarech (první případ, *nearly $17.3 billion*), nebyl by s tím zřejmě větší problém, pokud by to bylo pevně stanovené pravidlo, které by se konzistentně dodržovalo v celém treebanku.

- Nekonzistence kolem *more than* plynou z otázky, jak se vypořádat s předložkou *than*. V PCEDT je obecně pro toto spojení (tedy ne jen ve spojení s číselnými výrazy) používáno – bohužel nekonzistentně – těchto pět variant (viz obrázek 6.18):
  1. V nejčastějším případě je *more* chápáno jako substantivní, řídící uzel (v protikladu k adjektivní číslovce závislé na jiném substantivu), a to byl také preferovaný způsob anotace, kdykoli to bylo možné.
     ↪ *Příklad: „More than 2,100 others escaped to West Germany through Hungary over the Weekend.“* (strom č. 1 na obrázku 6.18). Pokud odstraníme rozvíjející větné členy, získáme parafrázi *More escaped to West Germany* (analogicky k *many/most/some/etc. escaped*), zatímco při uspořádání, kdy by *more than 2,100* bylo rozvíjejícím členem *others*, by redukce rozvíjejících větných členů vedla k parafrázi *Others escaped to West Germany*. Rozdíl v obou anotacích odráží dvě

[45] Možná s výjimkou složitější případu *more than*, který rozebereme v následujícím bodě.



Obrázek 6.17: Čtyři možná zavěšení „univerzálního atributu“. Přestože se čtyři modifikátory *nearly, at least, more than* a *about* chovají stejně (upřesňují částku, která za nimi následuje), v PCEDT bývají zavěšeny všemi čtyřmi možnými způsoby.

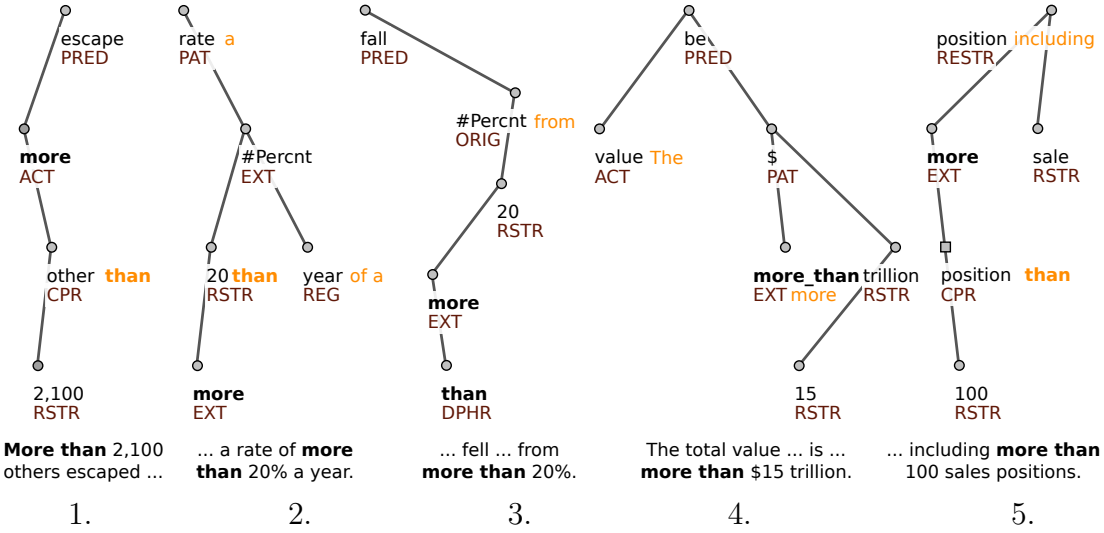
možné perspektivy pohledu na tutéž událost.

V uvažovaném preferovaném případě je tedy t-uzel *more* řídícím členem celé fráze a *than* je anotováno jako pomocné slovo, tedy bez vlastního uzlu na t-rovině (pouze odkazované pomocí `aux.rf` z podřízeného uzlu. Přestože z našeho úhlu pohledu to nebylo nejšťastnější rozhodnutí (neboť to stále považujeme za rozvití, které přináleží jinému výrazu), je výsledná anotace z pohledu tektogramatiky nejméně problematická ze všech pěti případů.

V PCEDT se vyskytuje 721×.

2. Druhá možnost je chápat *more* skutečně jakožto modifikátor a zavěsit ho jako závislý uzel – pak ovšem neexistuje (z pohledu tektogramatiky) moc vhodný uzel, kterému přiřadit *than*. Řeší se to přiřazením řídícímu uzlu. Tato anotace byla použita například ve větě *„All have a fiveyear earnings growth rate of more than 20% a year.“*, která se ničím nevymyká.

   V PCEDT se vyskytuje 13×.

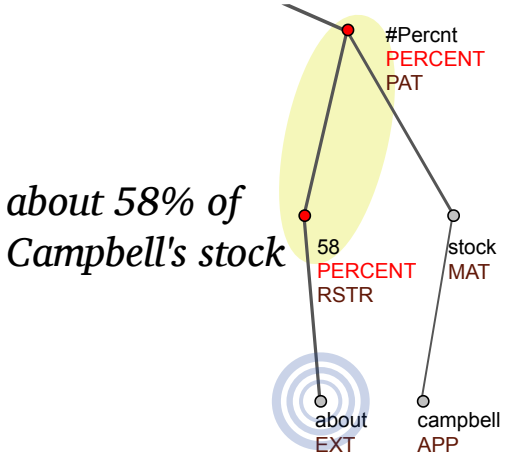


Obrázek 6.18: Nekonzistentní zachycení fráze *more than* v PCEDT. 1. *more* je uzel řídící a je rozvíjeno „množstvím“; 2. *more* je uzel závislý a rozvíjí množství spojené s pomocným *than*; 3. *more* rozvíjí množství, *than* je uzel DPHR; 4. *more_than* je složené `t_lemma`; 5. do stromu je vložen zkopírovaný uzel.

3. Jiný způsob, jak se vypořádat se zavěšením *more* pod výraz, (z nich také nejčastější), je zavedení samostatného uzlu DPHR pro *than*. To by ovšem správně nemělo nikdy nastat.
   V PCEDT se vyskytuje 152×.
4. Další způsob, jakým se anotátoři vyhnuli neintuitivnímu zavěšování „nad výraz“, je složené `t_lemma` *more_than*.
   V PCEDT se vyskytuje 16×.
5. Narazili jsme také na případy, kde byl zkopírován uzel v případech, kdy to neodpovídá anotačním instrukcím.
   ↪ *„…including more than 100 sales positions.“* nemá být anotováno jako *„more positions than 100 positions“*.

Mezi těmito případy po významové stránce však nevede ostrá hranice (pokud se vůbec dá nastavit) a bylo by proto vhodné je sjednotit.

- Zmínili jsme, že jsme převzali BBN entity tak, jak je do t-roviny PCEDT zanesl Josef Toman. I tato část vykazuje v některých případech vady, zejména pokud docházelo k dělení či slučování t-uzlů, ale nejen tam, viz obrázek 6.19.



Obrázek 6.19: Ukázka chybného převedení BBN entit na t-rovinu PCEDT. Slovo *about* (zvýrazněno) je původně správně součástí anotace, ovšem po přenesení anotací do PCEDT již součástí entity není.

6.2.4 Možnosti do budoucna

Jak jsme už uvedli, rádi bychom na základě těchto testů konzistence anotací přispěli k opravám (přinejmenším) treebanku PCEDT, který vzniká v našem ústavu.

Dosud jsme využívali BBN entity tak, jak je pro interní užití do PCEDT vložil Josef Toman. Nezachoval při tom ovšem rozsah jednotlivých entit: jsou pouze vyznačeny uzly, které jsou entitou zasaženy. (Proto jsou všechny entity, které z takovýchto dat extrahujeme, z definice souvislé.) V budoucnu tedy bude potřeba nejprve provést revizi, zachytit i hranice mezi případnými dvěma entitami stejného typu, které spolu sousedí (což je výjimečný případ, se kterým – pokud nastává – zacházíme chybně), a opravit další drobné chyby zmíněné v předchozí sekci.

V navazujícím projektu bychom rádi využili dvojjazyčnosti paralelního treebanku PCEDT. Chtěli bychom anglické entity projektovat na českou stranu a kvůli případným chybám si opět pomoci automatickou identifikací stejným způsobem na české straně. Získané pojmenované entity bychom rádi porovnali s automatickým značkovačem pojmenovaných entit pro češtinu, mj. i z teoretického hlediska (např. které pojmenované entity se do češtiny překládají tak, že přestanou být pojmenovanou entitou).

Aplikovali jsme algoritmus ze sekce 6.1.6 na (pojmenované) entity v angličtině. Při cross-validaci jsme dosáhli vyrovnaných výsledků mezi precision a recall, celkové F_1 score je mírně nižší než pro PDT (67 vs 73 %). Objevili jsme některé systematické chyby v Penn Treebanku, v anotaci entit od BBN, v jejich převodu do PCEDT i v samotném PCEDT. Máme v úmyslu se podílet na jejich opravě.

6.3 ČNK a VALLEX

V sekci 5.2 jsme popsali propojení starší verze VALLEXu 1.0 s novější verzí 2.5. Jak jsme zmínili, v rámci projektu VALEVAL byly navázány výskyty sloves ze skutečných textů ČNK na lexikální jednotky valenčního slovníku VALLEX 1.0. Propojení starší a novější verze VALLEXu jsme (vedle vyzkoušení metody propojování slovníků na základě valenčních rámců jednotek na jednodušším vstupu) prováděli právě za účelem navázání sloves z VALEVALu na nový VALLEX 2.5.

Nyní popíšeme, jak jsme tato data zpracovali a využili a jaký je veřejnosti zpřístupněný výsledek.

Pročištěná data VALEVALu po korekci neshod anotátorů (tzv. „golden VALEVAL“) obsahovala přes 500 LU a asi 8 000 anotovaných vět (plus předchozí kontext). Použili jsme ještě další data, která vznikla na konci projektu VALEVAL a která už do gold dat zahrnuta nebyla. Z těchto paralelních anotací jsme vybrali pouze takové věty, kde se všichni tři anotátoři zcela shodli na přiřazené LU. Získali jsme tak 9 465 vět pro 204 slovesa (a pro jejich 606 LU).

Některé zjevné chyby anotátorů, na něž jsme v průběhu práce narazili, stejně jako další drobné chyby jsme opravili. Také v případě 22 sloves, o nichž už jsme věděli, že se jejich LU od verze 1.0 k verzi 2.0 rozdělily, jsme provedli ruční „doanotaci“ a v každém sporném případě rozhodli, o kterého z následovníků původní LU se jedná.

Webová prezentace VALEVALu ve VALLEXu 2.6

Poté již bylo možné data hromadně zpracovat. Výstupem je jednak HTML rozhraní VALLEXu („pro lidi“) a jednak XML soubory („pro stroje“).

Pro každé sloveso[46] jsme vyrobili jeden HTML soubor se všemi větami, rozdělenými podle jednotlivých LU (viz obrázek 6.20). V každé větě je vyznačen výskyt dotyčného slovesa, který zároveň funguje jako odkaz do VALLEXu na použitou LU, dále kontext předcházejících tří vět, číslo věty ve VALLEXu a identifikátor dokumentu v ČNK.

[46] Připomeňme, že ve VALLEXu 1.0 nebyly vidové dvojice sdruženy do lexémů a příkladové věty jsou tudíž přiřazeny jednotlivým slovesným lemmatům.

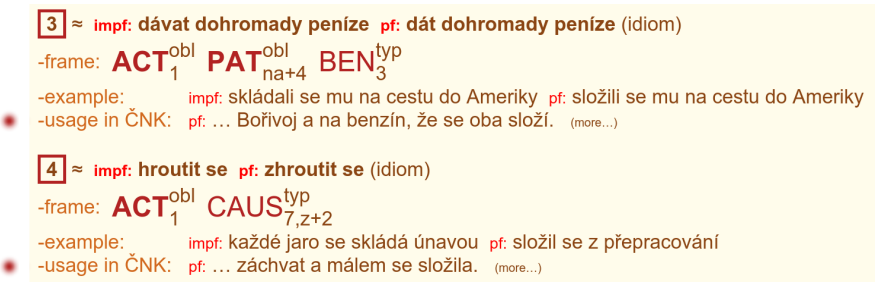
Obrázek 6.20: Webová stránka VALEVALu se všemi větami pro sloveso *složit se*. Lexikální jednotky 1 a 2 jsou omezeny jen na nedokonavé *skládat se*^{impf}, proto se zde nevyskytují, LU 3 má (jak je vidět na obrázku 6.21) význam „dát dohromady peníze“, LU 4 má význam „zhroutit se“.

Krátký úryvek z první věty každé lexikální jednotky jsme vložili též na stránky VALLEXu (viz obrázek 6.21) pod položku „usage in ČNK“. Odtud je možné v novém okně otevřít stránku se všemi větami na místě té LU, jejíž další výskyty nás zajímají.

V horním menu webového slovníku přibyla záložka VALEVAL, která umožňuje procházet veškerá slovesa, která dostala v projektu svou anotaci.

Datový formát VALEVALu

Anotované věty VALEVALu jsme také zpřístupnili v jednoduchém XML formátu. Je rozdělen podle jednotlivých sloves, v rámci slovesa následují jednotlivé výskyty v textu. Každý takový případ obsahuje správně přiřazené číslo LU a čtyři věty: tři kvůli kontextu, čtvrtou s vyznačeným slovesem:



Obrázek 6.21: Výřez z webového rozhraní VALLEXu 2.6, kde jsou u slovesa *složit se* doplněny odkazy do VALEVALu (vyznačeny tečkou vlevo). V tomto seznamu lexikálních jednotek uvádíme jen krátký výsek jediné věty, odkaz však vede na seznam všech vět pro danou LU (obrázek 6.20).

```
<body>
  <verb lemma='brát' frames='10'>
    ...
    <occurence number='54' frame='2'>
      <sentence>Nebraňte se tolik svému osudu!"</sentence>
      <sentence>Aidan se odvrátila.</sentence>
      <sentence>Jeho řeči ji dráždily.</sentence>
      <sentence is_here='1'><word>Braly</word> jí naději,
                            a toho se děsila.</sentence>
    </occurence>
    ...
  </verb>
  ...
</body>
```

Zároveň jsme pro potřeby práce s VALEVALem připravili zjednodušenou verzi VALLEXu, která v XML formátu SAMR, kterému je snazší porozumět a dál ho zpracovávat, obsahuje pouze vybraných 204 sloves, kterých se anotace VALEVALu týká.

Oba XML soubory i k nim příslušná DTD jsou k dispozici ke stažení na `http://ufal.mff.cuni.cz/legacy/vallex/2.6/doc/valeval.html`.

Popsali jsme zde navázání lexikálních jednotek 204 sloves VALLEXu 2.5 na téměř 10 000 vět z ČNK. Přiřazení správných LU bylo provedeno ručně při para-

lelní anotaci, následné převedení původních LU na nové automaticky (sekce 5.2). Výstupem je rozšířené HTML rozhraní slovníku VALLEX a veškerá informace uložená v XML formátu, což bylo vydáno jako VALLEX verze 2.6.

6.4 PDT/PCEDT a VALLEX

Stejně tak, jako nám propojení VALLEXu 1.0 a 2.5 (sekce 5.2) umožnilo přidání korpusových dokladů do VALLEXu 2.5 (předešlá sekce 6.3), tak i propojení VALLEXu a PDT-Vallexu (sekce 5.1) nám přináší navázání slovníku na korpusy PDT a PCEDT, čemuž věnujeme tuto sekci.

Zatímco v předchozí sekci jsme získali téměř 10 000 příkladů pro zhruba 200 sloves, spojením s PDT-Vallexem jsme získali 75 240 příkladů (31 422 vět z PDT a 43 818 z české části PCEDT) pro 2 302 slovesných lemmat.[47]

Příklady přidáváme pouze k těm lexikálním jednotkám, které přesáhly stanovenou hranici pro spojení s valenčním rámcem z PDT-Vallexu (viz sekci 5.1) a mají tedy přiřazen svůj ekvivalent z druhého slovníku jak na webu, tak v XML datech. Dále přidáváme příklady také k takovým slovesům, která mají ve VALLEXu i v PDT-Vallexu jen jedinou jednotku. Předpokládáme, že i když je valenční rámec zapsán v obou slovnících natolik odlišně, že se jej nepodařilo spárovat, je (i vzhledem k snaze o úplnost jednotek ve VALLEXu) velmi pravděpodobné, že pro dané sloveso skutečně existuje jednotka jediná, a proto je můžeme automaticky spojit.

Další postup vedoucí k zpřístupnění korpusových příkladů u slovníkových jednotek je obdobný jako u VALEVALu:

Samostatný XML soubor má pro každé slovesné lemma a pro každou jeho spárovanou LU uveden seznam všech výskytů dané LU v PDT i PCEDT ve formě `id` uzlů z t-roviny a pro snazší práci je zde připojena také celá příslušná věta v povrchové podobě. Pro úplnost je pro LU vždy uvedeno také ID odpovídajícího rámce z PDT-Vallexu (v případě, že jedné LU z VALLEXu odpovídá více rámců, jsou podle nich rozděleny i příklady do jednotlivých bloků).

HTML stránka pro každé dotčené sloveso má obdobný tvar jako na obrázku 6.20 – pro jednotlivé lexikální jednotky uvádí příslušné korpusové příklady. V budoucnu přibyde i odkaz na vyhledání a zobrazení tektogramatických stromů jednotlivých vět (až to bude umožňovat PML Tree Query service).[48] Webové roz-

[47] Stejně jako VALLEX 1.0, ani PDT-Vallex nesdružuje vidové dvojice, proto příklady z VALEVALu, PDT i PCEDT jsou ve VALLEXu přiřazovány jednotlivým lemmatům, nikoli lexémům.

[48] `https://lindat.mff.cuni.cz/services/pmltq`

hraní VALLEXu je doplněno o odkazy na tyto stránky a vybrané příklady z nich, jak je vidět na obrázku 6.22.

Obrázek 6.22: Ukázka ze slovesa *táhnout* ve VALLEXu s novými příklady. Přidané příklady z PDT a PCEDT („usage in PDT/PCEDT“) jsou zvýrazněny značkou vlevo. Menší oranžové značky ukazují rozšíření VALLEXu o odkazy do PDT-Vallexu (sekce 5.1 VALLEX a PDT-Vallex) a příklady z VALEVALu („usage in ČNK“, sekce 6.3 ČNK a VALLEX).

Popsaná data budou uvolněna do konce roku 2015 jako součást VALLEXu 3.0 na adrese `http://ufal.mff.cuni.cz/vallex/3.0`.

7

Vnitřní propojování jednotek slovníku

Už s prvními slovníky vznikala potřeba odkazovat se od jednoho hesla na jiné – někdy bylo cílem poukázat na vzdálenější souvislosti mezi dvěma hesly, jindy byla hesla (téměř) totožná a šlo o ekonomický zápis patřičných informací, což dále pomáhá předejít nekonzistentním informacím. Ať už se používá v češtině tradičního slova „viz“ nebo jen značky (↑, →), odkaz byl v klasických tištěných slovnících nutně omezený. Obvykle totiž odkazoval pouze na hlavní heslo (někdy i v případě, že se týkal jen některé jeho části). Nespecifikoval, o jaký druh vztahu se jedná (výjimkou je například × – značka pro antonymum). Odkazem bylo obtížné dosáhnout složitějšího propojení (např. sítě vztahů či hierarchie), neboť základní uspořádání hesel mohlo být jen jedno – typicky lineární, abecedně řazené – a odkazy jen málo rozšiřují možnosti, nemá-li být slovník přehlcený nepřehlednou záplavou odkazů.

Také slučování slovníkových hesel do jakýchkoli skupin funguje v tištěném slovníku omezeně, totiž pouze jednosměrně. Běžně se všem slovům připisuje ve slovníku např. slovní druh, máme tedy u každého slovesa napsáno „v“, ale nemůže to sloužit k získání představy o všech slovesech.

7.1 Stávající vnitřní propojení moderních slovníků

Dnes jsou možnosti mnohem bohatší a je dobře, že se také hojně využívají. Jako příklady rozmanitosti zachycovaných vztahů mezi jednotlivými slovníkovými jednotkami uveďme WordNet, FrameNet a VALLEX.

WordNet

WordNet jsme podrobně popsali v sekci 4.5. Zde zdůrazníme hlavní rysy z hlediska propojenosti jeho jednotek. WordNet není slovník v klasickém smyslu, je to lexikální ontologie, jejímž převodem do tištěné podoby by se do značné míry ztratil její smysl, neboť WordNet nemá zřetelné lineární uspořádání. Tvoří síť tzv.

synsetů, tedy skupin výrazů zhruba synonymních, které jsou navzájem propojené celou řadou sémantických vztahů. Jedná se tedy o více než jen o propojení dvojic: jsou zde celé skupiny výrazů. Mezi těmito synsety jsou zachyceny vztahy hyponymie, hyperonymie, holonymie, meronymie, antonymie apod. Ve WordNetu jsou tedy „přes sebe“ jedna hierarchická, téměř stromová struktura synsetů a mnoho dalších (nesouvislých) grafů mezi týmiž synsety.

FrameNet

Taktéž o FrameNetu jsme již psali v sekci 4.6, zde se na něj podíváme znovu z pohledu vztahů mezi jeho jednotkami.

Základní jednotkou je jistě sémantický rámec (semantic frame, SF), který (částečně podobně jako synset ve WordNetu) „obsahuje“ lexikální jednotky (je jimi evokován). Tyto SF jsou mezi sebou pospojovány pomocí osmi relací (z nichž sedm si můžeme připomenout na obrázku 4.3 na straně 44). Tyto relace tvoří ve FrameNetu osm různých (nesouvislých) grafů.

Topologicky (neboli z hlediska provázání) tedy vypadají FrameNet i WordNet podobně: graf s různě ohodnocenými orientovanými hranami (relace), vrcholy obsahují množinu lemmat (ve WordNetu tvoří nejbohatší hierarchii substantiva, ve FrameNetu slovesa).

Ve FrameNetu však existují paralelně se zmíněnými vztahy mezi SF ještě jemnější vztahy: Elementy dvou propojených sémantických rámců jsou spolu také propojeny, pokud mají v druhém SF svůj ekvivalent (viz obrázek 4.4 na straně 45).

VALLEX

VALLEX 1.0 ve svém nativním formátu a také VALLEX 2.6 ve formátu SAMR obsahují odkazy na vidové protějšky i ortografické varianty jak pro celá slovníková hesla (odpovídající jednotlivým lemmatům), tak pro lexikální jednotky v nich. (Vznik formátu SAMR a odkazy v něm jsme popsali v sekci 4.13.)

Také můžeme říci, že téměř polovina všech LU (45 %) je propojena do 22 skupin, neboli je součástí jedné z 22 sémantických tříd. To je případ kategorizace, která by nebyla moc užitečná, pokud bychom neměli možnost získat všechny LU v určené třídě.

Ve VALLEXu je rovněž anotovaná reciprocita a kontrola, jednotlivým LU je přiřazován typ této vlastnosti. Tyto jednotlivé typy reciprocity a kontroly tudíž tvoří další skupiny vzájemně propojených lexikálních jednotek.

SemLex

Ve slovníku SemLex (sekce 4.9) jsou některé dvojice synonymních víceslovných výrazů propojeny odkazem.[1] Za většinu z nich vděčí slovník svým primárním zdrojům, z kterých byl sestaven, zejména Českému WordNetu, v pročištěném slovníku jich zůstalo 180. Anotátoři měli také možnost odkazy na synonyma přidávat, na to ovšem museli vědět, že ve slovníku synonymní heslo existuje.

7.2 Probíhající vnitřní propojování vybraných slovníků

V této části se soustředíme na probíhající změny v oblasti propojování jednotek uvnitř slovníku ve dvou slovnících, se kterými jsme v této práci nejvíce pracovali: ve valenčním slovníku VALLEX a ve slovníku víceslovných výrazů SemLex. V případě VALLEXu probereme propojení členů diatezí a alternací, pro SemLex popíšeme jednoduché propojování „příbuzných“ pojmů a zanořených výrazů.

7.2.1 VALLEX 3.0

Ve VALLEXu je snaha sloučit slovesa se stejným syntaktickým chováním, tedy například vidové protějšky, ale také členy diatezí. To vede k úspornějšímu popisu stejně jako ke kontrole konzistence. V tzv. rozvinutém tvaru slovníku jsou pak členy diatezí vzájemně propojeny.

Diateze[2]

Diateze jsou vztahy mezi povrchově-syntaktickými konstrukcemi v rámci jedné lexikální jednotky, jejichž odlišnosti se projevují různými vyjádřeními slovesných doplnění (v důsledku odlišného usouvztažnění situačních participantů a větněčlenských pozic). Rozlišuje se nepříznakový člen (aktivní rod slovesný) a příznakové členy. Diateze patří mezi tzv. *gramatikalizované alternace* stejně jako reciprocita (str. 4.2), která je zachycovaná již ve VALLEXu 2.5.

↪ *Příklad:* Typickým příkladem příznakového členu je pasivum a rozdíl mezi aktivním a pasivním členem diateze ukážeme na následujících větách pro sloveso *rozvádět*$^{V}_{I}$:

[1] Přesně řečeno, jako synonymum je uvedena základní forma (BASIC_FORM) synonymního hesla. Jak jsme ale již řekli v úvodu ke slovníkům, ve slovníku nejsou homonymní víceslovné výrazy, takže to je dostatečné (poznámka 4 na straně 22).

[2] Autor se podílí na technickém řešení.

„Společnost$_{\mathsf{ACT-nom}}$ *rozvedla po budově*$_{\mathsf{LOC}}$ *Internet*$_{\mathsf{PAT-acc}}$*.“*
„Po budově$_{\mathsf{LOC}}$ *byl (společností*$_{\mathsf{ACT-inst}}$*) rozveden Internet*$_{\mathsf{PAT-nom}}$*.“*.

Ve VALLEXu 3.0, který vyjde v letošním roce, je zachyceno následujících pět typů diatezí (přičemž poslední typ byl již ve verzi 2.5):

- **pasivní diateze** (*Konečně byl kytovec uloven.*),
- **rezultativní diateze s *být*** (*Kurs Nautilu je již nastaven.*) a **s *mít*** (*Nemo má loď přichystánu k vyplutí.*),
- **recipientní (pasivní) diateze** (*Harpunář Ned Land dostal nařízeno ihned vystřelit.*),
- **dispoziční diateze** (*Mezi korýši se jim plulo snadno.*),
- **deagentní diateze** (*Kytovci se pro chutné maso loví od nepaměti.*).

Lopatková, Žabokrtský a Skwarska (2006); Kettnerová, Lopatková a Bejček (2012b) a Vernerová, Kettnerová a Lopatková (2014) navrhují tzv. *alternační model slovníku* rozdělující slovník na datovou komponentu (se kterou jsme dosud pracovali výhradně) a gramatickou komponentu. Diateze se pak zachycují v datové komponentě (atribut „diat“) u lexikální jednotky nepříznakového členu jako výčet typů diatezí, do kterých může daná LU vstupovat. V gramatické komponentě slovníku jsou pak uložena pravidla, která umožní z rámce odpovídajícího nepříznakovému užití LU v datové komponentě derivovat zápisy korespondující s ostatními přípustnými příznakovými užitími, jak navrhuje též Urešová (2011a).

Pokud pravidla aplikujeme, získáme tzv. *rozvinutý* tvar slovníku, ve kterém budeme mít pro jednu lexikální jednotku zaznamenáno více různých valenčních rámců (lišících se jen morfematickými formami), které mezi sebou budou vzájemně provázány relací diateze, ohodnocenou jejím typem. Minimální i rozvinutá varianta s provázanými valenčními rámci budou součástí vydání VALLEX 3.0.

Lexikalizované alternace ve VALLEXu[3]

Lexikalizovaná alternace je vztah mezi dvěma různě strukturovanými valenčními rámci, tedy mezi dvěma lexikálními jednotkami. Tyto dvě jednotky jsou – na rozdíl od diatezí – zachyceny ve VALLEXu odděleně. Jejich vzájemný vztah je však zachycen v podobě odkazu z jedné LU k druhé a naopak.

↪ *Příklad:* Uvedeme lexikálně-sémantickou konverzi materiálu a produktu ve dvojici vět
„Chlapec$_{\mathsf{ACT}}$ *na podlaze*$_{\mathsf{LOC}}$ *skládal kostky*$_{\mathsf{PAT}}$ *do pyramidy*$_{\mathsf{EFF}}$*.“* a
„Chlapec$_{\mathsf{ACT}}$ *na podlaze*$_{\mathsf{LOC}}$ *skládal z kostek*$_{\mathsf{ORIG}}$ *pyramidu*$_{\mathsf{PAT}}$*.“*.

VALLEX 3.0 obsahuje ruční anotace tří typů lexikalizovaných alternací, jak je rozlišuje Kettnerová a Lopatková (2013) a Kettnerová (2014, zde také detailní informace o alternacích):

[3] Autor řeší většinu technické části.

- **lexikálně-sémantické konverze** (dvojice *„Kytovec*$_{\mathsf{ACT}}$ *prorazil lodní trup*$_{\mathsf{PAT}}$*.“* a *„Kytovec*$_{\mathsf{ACT}}$ *prorazil do lodního trupu*$_{\mathsf{DIR3}}$ *díru*$_{\mathsf{PAT}}$*.“*),
- **dvojí strukturní realizace téhož situačního participantu** (dvojice *„Ned*$_{\mathsf{ACT}}$ *namířil harpunou*$_{\mathsf{MEANS}}$ *proti útočníkovi*$_{\mathsf{DIR3}}$*.“* a *„Ned*$_{\mathsf{ACT}}$ *namířil harpunu*$_{\mathsf{PAT}}$ *proti útočníkovi*$_{\mathsf{DIR3}}$*.“*),
- **strukturní rozpad situačního participantu** (dvojice *„Prof. Arronax*$_{\mathsf{ACT}}$ *viděl,* $_{(}$*jak harpuna mizí ve vlnách*$_{)\mathsf{PAT}}$*.“* a *„Prof. Arronax*$_{\mathsf{ACT}}$ *viděl harpunu*$_{\mathsf{PAT}}$*,* $_{(}$*jak mizí ve vlnách*$_{)\mathsf{EFF}}$*.“*).

Tyto dvojice alternujících valenčních rámců jsou ve VALLEXu 3.0 zachyceny odděleně jako odlišné LU, jsou však spojeny jedním ze tří typů relace alternace (jsou to ve stejném pořadí `conv`, `multiple` a `split`). Relace spojující dvě LU je navíc ohodnocena přesnějším podtypem relace (v našich příkladech „Affected_object–Hole“, „Instrument“ a „Stimul“), jak navrhuje Kettnerová (2014) a s čímž počítá také návrh formátu VALLEXu (Žabokrtský, 2005).

Vznikne tak propojení dvojic lexikálních jednotek[4] uvnitř lexému.

V budoucnu bychom se rádi věnovali také zachycení a propojení prefigovaných jednotek. Tím bychom rozšířili dosud velmi úzké pojetí vidovosti ve VALLEXu, ve kterém se uvažují obvykle jen dvojice a nepočítá se se vztahem jako např. *jít* $\longleftrightarrow$ {*dojít, přijít, odejít, přejít, obejít, …*}.

7.2.2 SemLex

Vztahy v SemLexu[5]

V SemLexu (sekce 4.9) je vedle popsaných synonym také velké množství dvojic víceslovných výrazů, které jsou částečně synonymní, nebo spolu nějak jinak úzce souvisí, a obvykle se užívají velice podobně. Některé z nich typově spadají do kategorií probíraných na stránkách 152–156,[6] zejména zde najdeme eliptické vztahy a lexikální varianty, hyperonyma a meronyma, ale také zanořené výrazy,

[4] Jedna LU ovšem může vstupovat do více různých alternací, být proto členem různých propojených dvojic.

[5] Toto je práce zejména autora.

[6] Na zmíněných stranách jsme psali, že některé tyto dvojice výrazů anotátoři zkrátka označili společným heslem ze slovníku a některé dokonce takto anotovat chceme. Zde budeme mluvit o těch případech, kdy pro nějakou podobnou dvojici existují ve slovníku dvě samostatná slovníková hesla (například proto, že tam byla vložena už při prvotním sestavování SemLexu z jiných slovníků a anotátoři neřešili na otázku, zda vložit nové heslo, nebo použít podobné).

anonyma, deverbativa nebo drobnosti jako změny v psaní velkých písmen. Takto najdeme v SemLexu například

- hyperonyma (např. sousloví *udržitelný rozvoj* se souslovími *trvale udržitelný rozvoj* a *dlouhodobě udržitelný rozvoj*),
- meronyma (*státní správa* a její součást *ústřední orgán státní správy*),
- jejich kombinace (*branný a bezpečnostní výbor sněmovny – branně bezpečnostní výbor parlamentu – branný výbor parlamentu – branně-bezpečnostní výbor parlamentu – branný výbor – předseda branného výboru parlamentu – branný a bezpečnostní výbor*),
- zanořování (*fotbalová liga* a *první fotbalová liga*) – o zanořování viz níže,
- jiné těsné vztahy (*loutkové divadlo – loutkové divadelnictví – loutkové představení*).

Nejen že spolu tyto dvojice souvisí významově, ale často spolu sdílí i idiosynkratické vlastnosti a nepravidelnosti, zkrátka rysy, které z nich dělají víceslovný výraz.

Tyto „příbuzné“ víceslovné výrazy jsou mezi sebou (zatím ručně) propojeny. Jedná se o pilotní studii vztahů, v rámci níž bylo propojeno 168 výrazů, zatím bez upřesnění druhů těchto vztahů. Příklad takového propojení byl také vidět v sekci 4.9 SemLex na obrázku 4.6. Pro příští vydání slovníku je chceme vyhledat automaticky (právě na základě podobnosti zápisu) a ručně jen zkontrolovat.

Zde uvádíme již bez kategorizace některé další zajímavé případy:

- deficitní rozpočet – deficit státního rozpočtu
- světová válka – první světová válka – druhá světová válka
- volební kampaň – předvolební kampaň
- na míru – šitý na míru – ušít na míru
- sportovní gymnasta – sportovní gymnastika
- dlouhodobá úroková sazba – úroková sazba – krátkodobá úroková sazba
- stavební kámen – základní stavební kámen
- samosprávný celek – vyšší územní celek – vyšší samosprávný celek – vyšší územní samosprávný celek – vyšší územně samosprávný celek – vyšší územněprávní celek – vyšší územněsprávní celek – územněsprávní celek
- věci veřejné – věc veřejná – veřejná věc

Zanořené víceslovné výrazy[7]

Zde nejprve nastíníme představy o rozšíření slovníku SemLex a tím se dostaneme k zanořeným výrazům.

Během projektu Lexemann (sekce 6.1.3) jsme do slovníku SemLex vkládali převážně víceslovné lexie. Anotátoři měli povoleno vkládat (s patřičnou poznám-

[7] Autor řeší ve spolupráci s P. Straňákem.

kou) i víceslovné pojmenované entity (named entities, NE), pro projekt ovšem takové záznamy nebyly potřeba a anotátoři to také (s výjimkou hraničních případů mezi víceslovnou NE a víceslovnou lexií) nevyužívali.

Do budoucna ovšem plánujeme některé typy NE do SemLexu hromadně vložit, neboť slovník umožňuje zachytit další pragmatické informace, které se pro zpracování jazyka dají využít. Příkladem může být tzv. *grounding*, tedy ukotvení pojmu, například pomocí odkazu na příslušný článek Wikipedie (pro typ objekt), nebo doplněním GPS souřadnic (pro typ místo).

Když uvažujeme o *vnitřní struktuře*, o vztazích mezi jednotlivými členy těchto víceslovných výrazů (VV) ve slovníku – lexií i pojmenovaných entit – zjistíme, že zdaleka ne všude dokážeme najít závislostní vztah, který nacházíme mezi téměř všemi ostatními jednotkami ve větě. Jedná se o plynulou škálu od neoslabených závislostních vazeb na jedné straně (kde jsou kolokace, které za VV nepovažujeme, a některé víceslovné lexie) přes oslabenou závislost, pevnou nezávislostní strukturu až po něco, co označujeme jako linearizace hodnot z tabulky. Více vysvětlí obrázek 7.1.

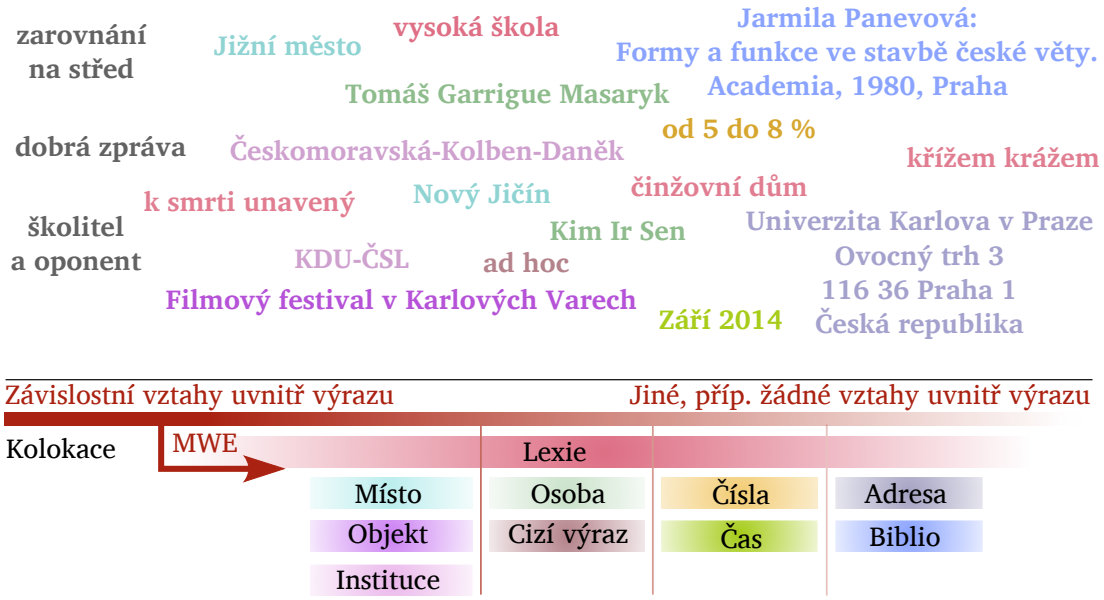
Vpravo na obrázku vidíme bibliografický záznam a adresu, v nichž hledat *závislostní* vztah mezi poštovním směrovacím číslem a městem nebo mezi rokem vydání a názvem publikace lze opravdu stěží. Přesto je ona vnitřní struktura velice pravidelná, jedná se však spíše o jakousi myšlenou tabulku, ve které jsou vyplněny některé řádky jako např. ulice, PSČ, číslo popisné a naopak nejsou vyplněny jiné, jako P. O. box či telefonní číslo. Podobná situace je například pro bibliografické záznamy. Tato fiktivní tabulka klíč–hodnota je linearizovaná do textové podoby (resp. obvykle pouze její hodnoty).

Pojmenované entity z pravé části škály do slovníku vkládat nechceme, nic jako „slovník všech adres“ nepotřebujeme. (V anotovaném korpusu navrhujeme celou takovou NE reprezentovat jediným uzlem jako je tomu v PDT od verze 2.5 a její vnitřní strukturu tam reprezentovat ve formě zmíněné tabulky. To však již vybočuje z tématu práce a zde se tím nebudeme dále zabývat.)

Vlevo naopak leží VV s poměrně zřejmou a standardní závislostní strukturou (*Filmový*$_{\mathsf{RSTR}}$ *festival v Karlových*$_{\mathsf{RSTR}}$ *Varech*$_{\mathsf{LOC}}$),[8] které navrhujeme do slovníku vkládat podobně jako lexie a dodat k nim „grounding“ a další pragmatické informace jako upřesnění podtypu (pro místo např. řeka, náměstí, stát, pro objekt např. kniha, chemikálie, zákon či jednotka).

Uprostřed na škále leží např. jména osob a cizí výrazy, jejichž součásti si uchovávají vztahy do určité míry závislostní. Lze argumentovat, že křestní jméno je

[8] Pochopitelně tím nechceme říct, že zde má slovo *Karlovy* ještě dnes stejný význam, jako třeba ve větě *„Rodiče potěšily Karlovy studijní výsledky.“*. Jedná se o VV, významy i vztahy jsou tudíž trochu posunuté.



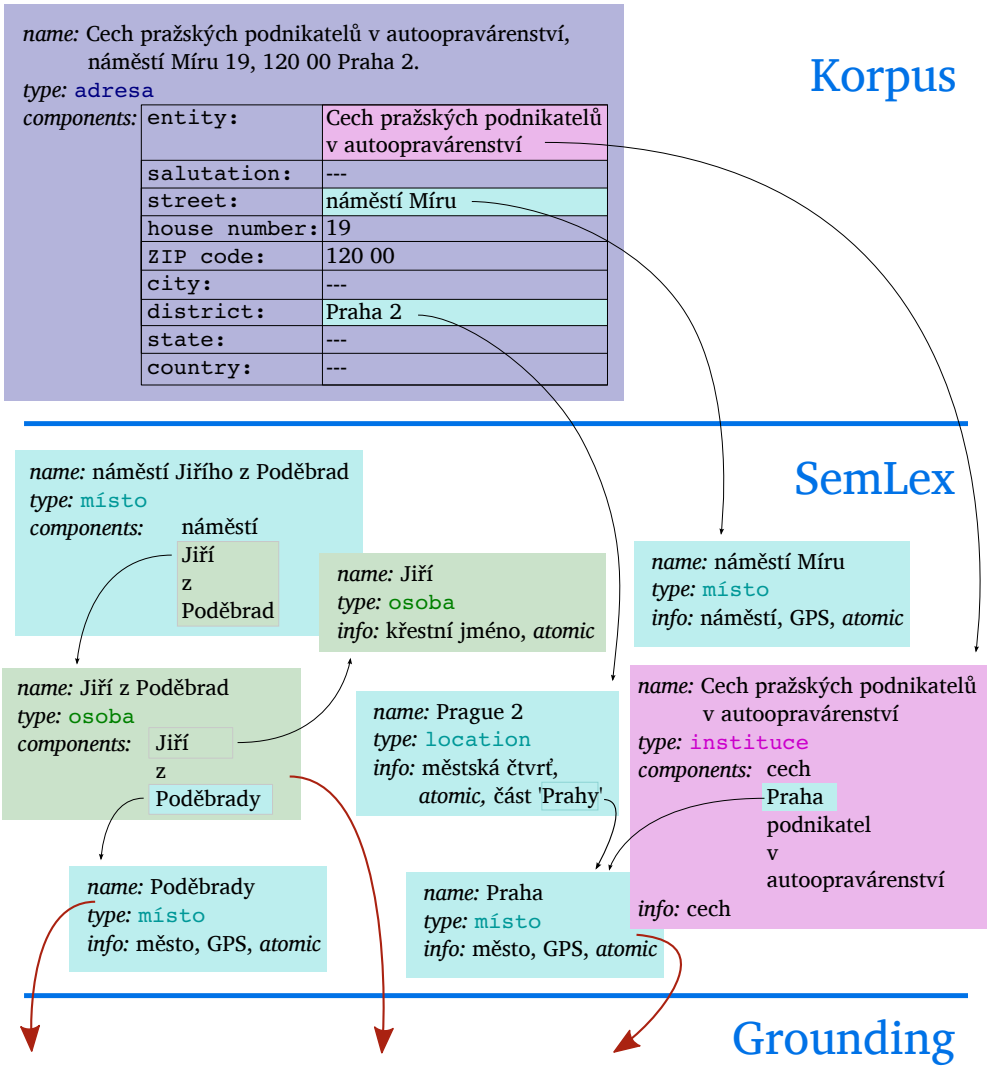
Obrázek 7.1: Škála vztahů uvnitř VV od ryze závislostních po pevnou nezávislostní strukturu. Vlevo vidíme běžné kolokace a některé VV, které mají pravidelnou závislostní strukturu (*Jižní* je atribut *města*, *dobrý* je atribut *zprávy*). S postupem vpravo se vztahy stávají nejasnějšími. (Je *Jičín* opravdu *nový*? Rozvíjí, tedy zpřesňuje měsíc letopočet, nebo letopočet měsíc?) Vpravo jsou zcela ustrnulé výrazy a spíše formální výčty hodnot nějakých atributů (jméno, ulice, PSČ, rok vydání, …).

dnes chápáno jako upřesnění příjmení (ačkoli historicky je to opačně), často to však není vůbec důležité. K tomu, abychom věděli, že je *Zhang Yimou* (někdy psaný jako *Yimou Zhang*) čínský režisér, nepotřebujeme rozkrýt strukturu jeho jména – a v případě vzdálených kultur to ani není snadné.

Tyto prostřední NE bychom rádi (například kvůli „groundingu“) vkládali do SemLexu, v případě jmen osob bychom ovšem známou vnitřní strukturu řešili tabulkou podobnou té pro adresy či biblio.

Součástí vnitřní struktury všech VV (lexií, „závislostních“ objektů i „tabulkových“ bibliografických záznamů) může být vnoření další (víceslovné) entity. Tyto součásti zachytíme jako odkazy na patřičné heslo ve slovníku, jako je to vidět na obrázku 7.2, u něhož budou vyplněny všechny další pragmatické informace.

Celý slovník tak bude vnitřně propojený, slovníková hesla se na sebe budou odkazovat (v některých případech bude kvůli „groundingu“ vhodné přidat i jednoslovná hesla).



Obrázek 7.2: Zachycení adresy v korpusu a náměstí v SemLexu s odkazy. Dlouhá adresa nahoře je v korpusu zachycena jako tabulka s odkazy z patřičných polí na VV do SemLexu. V SemLexu je navíc zachyceno *náměstí Jiřího z Poděbrad*, taktéž s odkazy na další (víceslovné) výrazy. Ty se mohou dále štěpit až na „atomické“ výrazy. Červené šipky symbolizují odkazy do externích zdrojů typu Wikipedie.

7.2.3 Závěr

Naším cílem nebylo podat ucelený obraz o vnitřním propojování slovníků. V práci pojednávající o propojování lexikografických zdrojů by však tato oblast neměla být zcela opominuta. Neboť vedle propojování slovníků mezi sebou a propojování

korpusových dat s těmito (propojenými) slovníky je i propojování a zahušťování informací uvnitř slovníku – stejně jako navazování na pragmatické informace o světě – významným směrem.

Nejprve jsme připomněli, jaká je současná situace ve slovnících, se kterými jsme v tomto textu pracovali. Potom jsme přešli k probíhajícím projektům nad VALLEXem a SemLexem.

Na příkladech těchto dvou slovníků jsme uvedli příklady jejich vnitřního provazování: (i) propojení různých povrchových realizací valenčního rámce pro diateze v rozvinutém tvaru slovníku VALLEX, (ii) propojení alternujících lexikálních jednotek tamtéž (iii) propojení příbuzných víceslovných výrazů v SemLexu a (iv) propojení víceslovných výrazů s výrazy v nich zanořenými tamtéž.

Závěr

Lexikální databáze, anotované korpusy, treebanky propojené se slovníky – všechny tyto jazykové zdroje se stále hojně využívají při počítačovém zpracování přirozeného jazyka. Ačkoli jich existuje velmi mnoho, jsou zde stále nejméně dva dobré důvody, proč se zajímat o jejich rozšiřování: Prvním důvodem je rozšíření pokrytí a zmnožení dostupných informací, což nastane pro každé propojení různých jazykových zdrojů. Druhým důvodem je situace pro jazyky, které dosud nejsou na jazykové zdroje bohaté. Pro ně je automatické propojení těch mála zdrojů, které existují, velkým přínosem, stejně jako navázání na nějaký „pivotní“ slovník v jiném jazyce, například anglický princetonský WordNet. Při tom se také zúročí metody, které byly vyvinuty pro větší jazyky z předchozího důvodu.

V této práci jsme po představení několika korpusů a treebanků a celé řady slovníků a lexikálních databází navrhli a popsali automatické metody, které lze využít pro propojování dat. Tyto metody jsme aplikovali na vybrané jazykové zdroje a otestovali jsme tak jejich využitelnost v praxi. Vedle nízké ceny aplikace *automatické* metody spočívá její výhoda také ve faktu, že je možné ji pouštět opakovaně na nové verze slovníků a udržovat tak aktualizované propojení, popřípadě nové verze slovníků okamžitě propojit s korpusem.

Formát dat

Studium formátů lexikografických zdrojů v této práci citovaných nás přesvědčilo, že dosud naprostá většina autorů slovníků spoléhá na svůj vlastní proprietární formát, který si vytvoří přesně na míru svým potřebám: nic v něm nechybí a neobsahuje nic nadbytečného.

Po seznámení se standardy pro slovníkové formáty máme za to, že debata o těchto univerzálních formátech v souvislosti s propojováním jazykových dat řeší zástupný problém. Důležitá a zcela zásadní je otázka, jaké informace slovník skýtá (zda jsou užitečné pro porovnávání, zda jsou k dispozici v obou slovnících v obsahově porovnatelné podobě, zda jich je dostatek). V jakém formátu jsou

tyto informace uloženy, je podle naší zkušenosti věc podružná, snadno řešitelná (i když někdy technicky pracná).

Podobně jako autoři řady slovníků jsme se i my pro účely propojování dvojice slovníku VALLEX a PDT-Vallex rozhodli setrvat u vlastních formátů, dosud pro oba slovníky používaných. Na mírnou úpravu slovníků do nově navrženého odvozeného formátu tak, aby se sobě slovníky více podobaly, bylo potřeba mnohem menší úsilí než na vytvoření zcela nového formátu na podkladu univerzálního frameworku. Ten by přitom oplátkou nepředstavoval přínos, neboť všechny potřebné jevy už nyní v používaných formátech zachytit umíme.

Propojování dat

O výhodách propojování jazykových dat jsme zde již psali několikrát. Konkrétním konečným cílem propojení slovníků může být ve výjimečných případech vytvoření jediného slovníku – nástupce obou propojovaných, přičemž původní slovníky „zaniknou“. Ve většině případů je však nutné oba slovníky zachovat – je tomu tak např. v projektu SemLink propojujícím anglické slovníky a lexikální databáze či v projektu EuroWordNet, kde byla vytvořena propojená síť různojazyčných WordNetů. Důvodem mohou být např. různá teoretická východiska a využití v různých projektech. Navrhli jsme tedy metodu, která je automatická (tudíž levná) a lze ji aplikovat opakovaně, jak se propojované jazykové zdroje budou vyvíjet, aniž bychom přišli o případné ruční opravy. Automatické propojování slovníku s textovými daty přináší zejména důležité dodatečné informace pro zpracování textu: minimálně desambiguaci jazykových jednotek a jejich obohacení o všechny další informace ve slovníku k heslům ukládané.

Zásadní praktická aplikace metody na propojování zdrojů je propojení dvojice **VALLEX ↔ PDT-Vallex**, kde jsme popisovanou automatickou metodou dosáhli úspěšnosti 77 % (F_1 score). Považujeme to za dobrý výsledek, mj. jsme výrazně překonali baseline 63 %. Tyto výsledky jsme však ještě upravili a za cenu nižšího pokrytí lexikálních jednotek jsme odstranili méně důvěryhodné odkazy, čímž jsme dosáhli přesnosti 91 % – tedy stejné precision ve srovnání s mezianotátorskou shodou při ručním mapování. Metoda je automatická a tudíž v budoucnu opakovaně použitelná pro nové verze slovníků.

Výstupem jsou dvojice odpovídajících si lexikálních jednotek/valenčních rámců. Odkazem do PDT-Vallexu jsme propojili 88 % lexémů z VALLEXu (těch, které mají svůj protějšek v PDT-Vallexu) a v nich 1 947 lexikálních jednotek získalo odkaz na valenční rámec v PDT-Vallexu. Pomocí propojení VALLEX ↔ PDT a VALLEX ↔ PCEDT navíc získaly lexikální jednotky z VALLEXu napojení na desítky tisíc anotovaných vět v obou pražských závislostních korpusech a díky

propojení VALLEX ↔ VALEVAL získaly pro zhruba 200 sloves téměř 10 000 příkladových vět z ČNK.

Stále běžící projekt propojující dvojici **VALLEX ↔ FrameNet** je jediným, který překračuje hranici češtiny a klade si za cíl propojení s anglickou databází. Takové propojení by také přineslo potřebnou sémantickou informaci do VALLEXu, proto je to téma, kterému se chceme dále věnovat. Poté již bude snadné propojit také dvojici PDT-Vallex ↔ FrameNet.

V druhé části práce jsme nejprve přinesli rozsáhlý popis některých vlastností víceslovných výrazů ve vztahu k tektogramatické rovině. Probírali jsme využití těchto vlastností ve slovníku víceslovných výrazů SemLex i překážky, které tyto vlastnosti kladou do cesty slovníku a zejména pak metodě na slovníku založené. Probíranou metodou propojující dvojici **SemLex ↔ ČNK** je automatická identifikace VV v korpusu ČNK.

Dosažená úspěšnost metody pracující na významové rovině jazyka se za současného stavu metody i této úrovně popisu nejeví jako zřetelně výhodnější než postupy na jiných méně abstraktních rovinách. Ovšem metodu je možné dále vylepšovat, návrhy zmíněné v diskusi hodláme uvést do praxe, neboť se domníváme, že zanedbání roviny jazykového významu dlouhodobě lepší výsledky nepřinese.

Naposledy zmíníme dvojici **slovník BBN entit ↔ PCEDT**, jejíž propojení v této práci funguje jako test konzistence anotace (od anotace Penn Treebanku přes anotaci BBN entit až po anotaci anglické části PCEDT). Nalezli jsme některé systematické chyby, jejichž vyřešením lze kvalitu anotace v PCEDT zlepšit.

Seznam obrázků

Seznam tabulek

Používaná terminologie

Slovníček pojmů

aktant je prominentní valenční doplnění (typicky objektové či subjektové povahy); ve FGP se rozlišuje pět aktantů označených funktory ACT, PAT, EFF, ADDR nebo ORIG

analytický predikát je konstrukce sémanticky vyprázdněného slovesa a predikativního substantiva, např. *podat žádost* či *mít nápad*

delimitace ↑*lexikálních jednotek* je lexikografický proces dělení ↑*lexému* na lexikograficky dobře vymezené a specifikované lexikální jednotky

efektivní rodič/potomek je v PDT symbolická relace v závislostním stromě mezi řídícím uzlem a uzlem, který ho modifikuje, která odstiňuje problémy s koordinací

↪ *Příklad:* V závislostním stromě ve stylu PDT je pro větu *„Vyhodili staré noviny a časopisy“* koordinační spojka *a* technickým rodičem uzlů *noviny* i *časopisy*, stejně jako je rodičem i jejich společného rozvití *staré*. Zato *efektivním* rodičem adjektiva *staré* jsou obě modifikovaná substantiva naráz a efektivním rodičem těchto substantiv je sloveso *vyhodit* – jako by zde žádná koordinační spojka ani nebyla.

Funkční generativní popis jazyka, FGP je jazyková teorie, jejíž součástí je teorie ↑*valence*, se kterou zde pracujeme, a která je teoretickým základem PDT (Sgall, 1967; Sgall et al., 1986; Panevová, 1974); jinde též FGD z anglického Functional Generative Description

heslo, slovníkové je základní položka slovníku; ve VALLEXu reprezentuje celý ↑*lexém*, ve formátu SAMR a v PDT-Vallexu reprezentuje ↑*lexikální jednotku*; v SemLexu je to víceslovná ↑*lexie*, případně výjimečně též víceslovná ↑*pojmenovaná entita*

instance jsou výskyty slova v textu, jako protiklad k ↑*typům*; terminologie pochází z komputační oblasti, v lexikografii by instanci ↑*lexikální jednotky* odpovídal *alolex*

iterativum zde používáme jako hodnotu atributu vid, a to pro opakující se děj, někdy též označované jako nedokonavé sloveso opakovací
↪ *Příklad: kopnout^{pf}, kopat^{impf}, kopávat^{iter}*

kontrola je určitý typ gramatické koreference u některých sloves, lepší vysvětlení i s příkladem je na str. 28

lemma je řetězec znaků, který reprezentuje slovníkovou položku, ↑*lexém*. V české tradici se obvykle jedná o nominativ singuláru pro substantiva, infinitiv pro slovesa apod. Pro reflexivní sloveso typu reflexivum tantum či derivované reflexivum tantum obsahuje *lemma* také reflexivní částici *„se/si“* (na rozdíl od ↑*m-lemmatu*).

lexém je abstraktní formálně-významová jednotka lexikonu; sdružuje množinu všech možných manifestací slovesa, neboli ↑*lexikálních forem* (celé paradigma) a množinu ↑*lexikálních jednotek*, které reprezentují jeho významové složky; celý *lexém* je reprezentován ↑*lemmatem* (případně několika lemmaty);
výraz *lexém* ve VALLEXu verze 2.0 a vyšší sdružuje také vidové protějšky a ortografické varianty, takže například slovesa *chytat^{impf}* a *chytit/chytnout^{pf}* sdílí jediný lexém a do jednotlivých ↑*lexikálních jednotek* jsou ↑*delimitována* společně;
naproti tomu v PDT-Vallexu, ve VALLEXu 1.0 i ve slovnících převedených do formátu SAMR jsou vidové protějšky či varianty reprezentovány vlastním lexémem a jednomu lexému pak vždy odpovídá jediné ↑*lemma* (uváděný příklad by byl pokryt třemi lexémy)

lexémový shluk, anglicky „lexeme cluster“, je množina lexémů, které mají stejné ↑*m-lemma* a liší se přítomností reflexivní částice, např. *„dovolit/dovolovat, dovolit/dovolovat se, dovolit/dovolovat si“*

lexie je slovo nebo slovní spojení v jednom konkrétním významu, v české jazykovědě má tedy obvykle stejný význam jako monosémický lexém a jako základní lexikální jednotka, viz sekci 6.1.2. V této práci budeme pojem *lexie* používat výhradně v části 6.1 pro víceslovnou lexii, jako protiklad k víceslovné pojmenované entitě.

lexikální forma, neboli tvar slova; všechny *lexikální formy* slova tvoří paradigma[9] a spolu se všemi ↑*lexikálními jednotkami* vytvářejí ↑*lexém*

lexikální jednotka, LU významová složka ↑*lexému*, slovo v daném významu (jedno ↑*lemma* tak může reprezentovat více lexikálních jednotek); v české lexikologii též *základní lexikální jednotka*, či ↑*lexie*

[9] Paradigma v tomto případě pro sloveso neobsahuje pouze jeden infinitiv, tvary pro tři osoby singuláru a plurálu, rozkazovací způsob, přechodníky, příčestí, ..., ale obsahuje také všechny vidové varianty a vyjmenované tvary pro každý z nich.

m-lemma zavádíme v souladu s Žabokrtským (2005) jako slovesnou část případně reflexivního slovesného ↑*lemmatu*, např. *umínit, divit*; v případě nereflexivních sloves pojmy ↑*lemma* a *m-lemma* splývají; (název značí morfologické lemma)

odpovídat si je relace označující dvě lexikální jednotky (případně i celá dvě slovesa), která vyjadřují v zásadě stejný význam. Cílem úlohy provazování slovníku je propojit jednotky, které si vzájemně *odpovídají*.

podstrom je souvislý podgraf stromu. Rozšiřujeme tak v této práci definici podstromu běžnou v teorii grafů. Nevyžadujeme, aby do něj náležely všechny uzly od nového kořene k listům, stačí, aby výsledkem byl strom (příp. strom s daným kořenem). Viz obrázek 6.2 na straně 142.

pojmenovaná entita, NE je v našem pojetí ustálený název (osob, institucí apod.), strukturovaný údaj (typu adresa), časový údaj, číselný rozsah nebo cizí slova, viz Pojmenované entity na str. 137

reciprocita je symetrické postavení dvou (a více) valenčních doplnění. Tradičním příkladem je reciprocita mezi ACT a PAT u slovesa *líbat*: *„Jan líbal Marii.* $\Longrightarrow$ *Jan a Marie se líbali.“*. Reciprocita se však může týkat i jiných funktorů: PAT+DIR3 *„Míchala do sebe cukr a žloutek.“* (4.2)

systémové uspořádání je ve FGP pořadí doplnění v samostatně stojící bezpříznakové větě, viz též str. 24

typy jsou abstraktní jazykové jednotky, které jsou popisovány ve slovníku, jako protiklad k ↑*instancím* v textu; terminologie pochází z komputační oblasti

`t_lemma` je řetězec znaků, obvykle slovo, reprezentující lexikální informaci příslušného autosémantického slova na t-rovině PDT

t-rovina je nejvyšší rovina jazykového popisu v korpusu PDT, vychází z tektogramatické roviny, tedy roviny jazykového významu ve FGP;

valence je schopnost plnovýznamového slova vázat určitý počet závislých členů a určovat jejich typ. Valence v teorii FGP je popsána v sekci 4.1 na straně 23.

víceslovný výraz je pro nás spojení alespoň dvou plnovýznamových slov, které tvoří dohromady jednotku s výrazně odlišným významem (anebo chováním), vykazují nekompozicionalitu, viz sekce Víceslovné lexie na str. 136. V této práci používáme také jako souhrnné označení pro víceslovné ↑*pojmenované entity* i víceslovné ↑*lexie*.

Jazykové zdroje

ČNK 3.1, Český národní korpus
FrameNet 4.6
PCEDT 3.5, Prague Czech-English Dependency Treebank

PDT 3.3, Prague Dependency Treebank
PDT-Vallex 4.3
Penn Treebank 3.2
PropBank 4.7.1
SemLex 4.9
SYN 2000 3.1
SemLink 4.8
VALLEX 4.2
VALEVAL 4.2.1
VerbNet 4.7.3
WordNet 4.5

Použité zkratky

BBN americká technologická firma, jejíž „BBN Entity Type Corpus“ obsahující ruční anotace (pojmenovaných) entit používáme

FE „frame element“ (česky element rámce), participant sémantického rámce (SF) ve FrameNetu

FGP zkratka názvu teorie ↑*Funkčního generativního popisu*

LU ↑*lexikální jednotka* („lexical unit“)

NE ↑*pojmenovaná entita* („named entity“), viz str. 137

PML Prague Markup Language, formát s vlastních schematem založený na XML, v němž je například uloženo PDT, PCEDT, ale i PDT-Vallex.

PML-TQ PML Tree Query, nástroj pro pokročilé komplexní vyhledávání nad daty v PML

SAMR zkratka formátu valenčního slovníku, 4.13

SF „semantic frame“ (česky sémantický rámec) ve FrameNetu zachycuje nějakou situaci, objekt či událost, viz sekce 4.6

SSČ je zkratka pro Slovník spisovné češtiny pro školu a veřejnost (Filipec et al., 2005).

TS používáme pro „tree structure“, tedy tektogramatický (či analytický) závislostní podstrom reprezentující víceslovný výraz, viz str. 141. Většina atributů uzlů do TS zahrnuta není, výjimku tvoří `t_lemma`.

VV používáme pro víceslovný výraz (zejména pro víceslovnou lexii), viz 6.1.2.

WSJ The Wall Street Journal, americký deník, základ Penn Treebanku i PCEDT

Literatura

Ajdukiewicz, K. (1935). Die syntaktische Konnexität. *Studia Philosophica*, I:1–27.

Bejček, E., Kettnerová, V. a Lopatková, M. (2010). Advanced searching in the valency lexicons using PML-TQ search engine. In Sojka, P., Horák, A., Kopeček, I. a Pala, K., editors, *Text, Speech and Dialogue. 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings*, volume 6231 of *Lecture Notes in Computer Science*, pages 51–58, Berlin / Heidelberg. Springer.

Bejček, E., Kettnerová, V. a Lopatková, M. (2014). Automatic mapping lexical resources: A lexical unit as the keystone. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. a Piperidis, S., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2826–2832, Reykjavík, Iceland. European Language Resources Association.

Bejček, E., Möllerová, P. a Straňák, P. (2006). The lexico-semantic annotation of PDT. In Sojka, P., Kopeček, I. a Pala, K., editors, *Lecture Notes in Computer Science, Proceedings of the 9th International Conference, TSD 2006*, volume 4188 of *Lecture Notes in Computer Science*, pages 21–28, Berlin / Heidelberg. Springer.

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J. a Žabokrtský, Z. (2012). Prague dependency treebank 2.5 – a revisited version of PDT 2.0. In Kay, M. a Boitet, C., editors, *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246, Mumbai, India. IIT Bombay, Coling 2012 Organizing Committee.

Bejček, E. a Straňák, P. (2010). Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Bejček, E., Straňák, P. a Hajič, J. (2009). Finalising multiword annotations in PDT. In *Proceedings of 8th Treebanks and Linguistic Theories Workshop (TLT)*, pages 17–25, Milano, Italy.

Bejček, E., Straňák, P. a Pecina, P. (2013). Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In

The 9th Workshop on Multiword Expressions (MWE 2013), pages 106–115, Atlanta, Georgia, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Bejček, E., Straňák, P. a Schlesinger, P. (2008). Annotation of multiword expressions in the Prague dependency treebank. In *IJCNLP 2008 Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 793–798, Hyderabad, India. Asian Federation of Natural Language Processing, International Institute of Information Technology.

Benešová, V., Lopatková, M. a Hrstková, K. (2008). Enhancing czech valency lexicon with semantic information from framenet: The case of communication verbs. In *ICGL 2008 Proceedings of the First International Conference on Global Interoperability for Language Resources*, pages 18–25, Hong Kong, China. City University of Hong Kong, City University of Hong Kong.

Bloomfield, L. (1933). *Language*. Holt, New York.

Bojar, O. (2009). *Exploiting Linguistic Data in Machine Translation*, volume 3 of *Studies in Computational and Theoretical Linguistics*. ÚFAL, Praha, Czechia.

Bojar, O., Semecký, J. a Benešová, V. (2005). VALEVAL: Testing VALLEX consistency and experimenting with word-frame disambiguation. *The Prague Bulletin of Mathematical Linguistics*, 83:5–17.

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M. a Tamchyna, A. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC2012*, Istanbul, Turkey. ELRA, European Language Resources Association.

Bonial, C., Bonn, J., Conger, K., Hwang, J. D. a Palmer, M. (2014). PropBank: Semantics of new predicate types. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. a Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bouamor, D., Semmar, N. a Zweigenbaum, P. (2012). Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 95–108, Mumbai, India. The COLING 2012 Organizing Committee.

Chomsky, N. (1957). *Syntactic Structures*. Mouton & Co.

Chomsky, N. (1970). Remarks on nominalization. In Jacobs, R. A. a Rosenbaum, P. S., editors, *Readings in English Transformational Grammar*, pages 184–221. Ginn, Boston.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Čermák, F., Červená, V., Churavý, M. a Machač, J. (1994). *Slovník české frazeologie a idiomatiky*. Academia, Praha.

de Lacalle, M. L., Laparra, E. a Rigau, G. (2014). Predicate Matrix: extending SemLink through WordNet mappings. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. a Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Declerck, T. (2006). SynAF: Towards a standard for syntactic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

Eckart, K. (2012). A standardized general framework for encoding and exchange of corpus annotations: The Linguistic Annotation Framework, LAF. In Jancsary, J., editor, *Proceedings of KONVENS 2012*, pages 506–515. ÖGAI. SFLR 2012 workshop.

Eryigit, G., Ilbay, T. a Can, O. A. (2011). Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (IWPT – 12th International Conference on Parsing Technologies)*, pages 45–55, Dublin, Ireland. Association for Computational Linguistics.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Filipec, J. (1994). Lexicology and lexicography: Development and state of the research. In Luelsdorff, P. A., editor, *The Prague School of Structural and Functional Linguistics*, pages 163–183, Amsterdam/Philadelphia. J. Benjamins.

Filipec, J., Daneš, F., Machač, J. a Mejstřík, V. (2005). *Slovník spisovné češtiny pro školu a veřejnost*. Academia, Praha. Zkráceně SSČ.

Filipec, J. a Čermák, F. (1985). *Česká lexikologie*. Number 20 in Studie a práce lingvistické. Academia.

Fillmore, C. J. (1968). The Case for Case. In Bach, E. a Harms, R. T., editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Fillmore, C. J. (1977). The case for case reopened. In Cole, P. a Sadock, J. M., editors, *Syntax and Semantics*, volume 8: Grammatical Relations, pages 59–81. Academic Press Inc.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M. a Soria, C. (2006). Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, MLRI '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z. a Žabokrtský, Z. (2012). Announcing Prague Czech-English dependency treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A., Štěpánek, J., Pajas, P. a Kárník, J. (2004). Anotace na analytické rovině. Návod pro anotátory. Technical Report TR-2004-23, ÚFAL/CKL MFF UK, Prague, Czech Republic.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V. a Pajas, P. (2003). PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In Nivre, J. a Hinrichs, E., editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Hajič, J., Vidová Hladká, B., Panevová, J., Hajičová, E., Sgall, P. a Pajas, P. (2001). Prague dependency treebank 1.0 (final production label). In *CDROM*. CAT: LDC2001T10., ISBN 1-58563-212-0.

Hana, J., Zeman, D., Hajič, J., Hanová, H., Hladká, B. a Jeřábek, E. (2005). Manual for morphological annotation, revision for the Prague dependency treebank 2.0. Technical Report TR-2005-27, ÚFAL MFF UK, Prague, Czech Republic.

Havránek, B., Bělič, J., Helcl, M. a Jedlička, A., editors (1989). *Slovník spisovného jazyka českého*. Academia, Praha. Zkráceně SSJČ.

Horák, A. (2008). *Computer Processing of Czech Syntax and Semantics*. Librix.eu, Brno, Czech Republic.

Jelínek, J., Těšitelová, M. a Bečka, J. V. (1961). *Frekvence slov, slovních druhů a tvarů v českém jazyce*. SPN, Praha.

Joshi, A. K. a Schabes, Y. (1991). Tree-adjoining grammars and lexicalized grammars. Technical Report MC-CIS-91-22, University of Pennsylvania, Department of Computer and Information Science. Paper 445.

Kempson, R. M. (1977). *Semantic Theory*. Cambridge University Press.

Kettnerová, V. (2014). *Lexikálně-sémantické konverze ve valenčním slovníku*. Karolinum, Prague, Czech Republic.

Kettnerová, V. a Lopatková, M. (2013). Lexikalizované alternace v češtině. *Linguistica Copernicana*, 9(1):31–64.

Kettnerová, V., Lopatková, M. a Bejček, E. (2012a). Mapping semantic information from framenet onto VALLEX. *The Prague Bulletin of Mathematical Linguistics*, 97:23–41.

Kettnerová, V., Lopatková, M. a Bejček, E. (2012b). The syntax-semantics interface of czech verbs in the valency lexicon. In Fjeld, R. V. a Torjusen, J. M., editors, *Proceedings of the 15th EURALEX International Congress*, pages 434–443, Oslo, Norway. Department of Linguistics and Scandinavian Studies, University of Oslo.

Laparra, E., Rigau, G. a Cuadros, M. (2010). Exploring the integration of WordNet and FrameNet. In *Proceedings of the 5th Global WordNet Conference (GWC'10)*, Mumbai (India).

Levin, B. (1993). *English Verb Classes and Alternation: A Preliminary Investigation*. PhD thesis, University of Chicago.

Lopatková, M. (2010). *Valency Lexicon of Czech Verbs: Towards Formal Description of Valency and Its Modeling in an Electronic Language Resource*. Habilitační práce, Charles University in Prague, Faculty of Mathematics and Physics, Prague.

Lopatková, M. a Panevová, J. (2006). Recent developments of the theory of valency in the light of the Prague dependency treebank. In *Insight into Slovak and Czech Corpus Linguistics*, pages 83–92. Veda Bratislava, Bratislava, Slovensko.

Lopatková, M., Žabokrtský, Z. a Kettnerová, V. (2008). *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha.

Lopatková, M., Žabokrtský, Z., Kettnerová, V., Skwarska, K., Bejček, E., Hrstková, K., Nová, M. a Tichý, M. (2007). VALLEX 2.5 - Valency lexicon of Czech verbs, version 2.5. Data.

Lopatková, M., Žabokrtský, Z. a Skwarska, K. (2006). Valency lexicon of Czech verbs: Alternation-based model. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1728–1733, Genova, Italy. ELRA, ELRA.

Loper, E., Yi, S. a Palmer, M. (2007). Combining lexical resources: Mapping between PropBank and VerbNet. In *In Proceedings of the 7th International Workshop on Computational Linguistics*.

Materna, J. a Pala, K. (2010). Using ontologies for semi-automatic linking VerbaLex with FrameNet. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M. a Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. a Grishman, R. (2004). *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, chapter The NomBank Project: An Interim Report.

Mikulová, M., Bejček, E., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Straňák, P., Ševčíková, M. a Žabokrtský, Z. (2013). Úpravy a doplňky pražského závislostního korpusu (od PDT 2.0 k PDT 3.0). Technical Report ÚFAL TR-2013-53.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K. a Žabokrtský, Z. (2006). Annotation on the tectogrammatical level in the Prague dependency treebank. Annotation manual. Technical Report TR-2006-30, ÚFAL MFF UK, Prague, Czech Republic.

Necsulescu, S., Bel, N., Padró, M., Marimon, M. a Revilla, E. (2011). Towards the automatic merging of language resources. In *First International Workshop on Lexical Resources: an ESSLLI 2011 Workshop*, pages 70–77, Ljubljana, Slovenia. ESSLLI.

Novák, V. a Žabokrtský, Z. (2007). Feature engineering in maximum spanning tree dependency parser. In Matoušek, V. a Mautner, P., editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 92–98, Berlin / Heidelberg. Springer.

Pajas, P. a Štěpánek, J. (2005). A generic XML-based format for structured linguistic annotation and its application to prague dependency treebank 2.0. Technical Report TR-2005-29, ÚFAL MFF UK.

Pajas, P. a Štěpánek, J. (2008). Recent advances in a feature-rich framework for treebank annotation. In Scott, D. a Uszkoreit, H., editors, *The 22nd International Conference on Computational Linguistics - Proceedings of the Conference*, volume 2, pages 673–680, Manchester, UK. The Coling 2008 Organizing Committee.

Pajas, P. a Štěpánek, J. (2009). System for querying syntactically annotated corpora. In Lee, G. a im Walde, S. S., editors, *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, Singapore. Association for Computational Linguistics.

Pal, S., Chakraborty, T. a Bandyopadhyay, S. (2011). Handling multiword expressions in phrase-based statistical machine translation. *Machine Translation Summit XIII*, pages 215–224.

Pala, K., Čapek, T., Zajíčková, B., Bartůšková, D., Kulková, K., Hlaváčková, D., Hoffmannová, P., Bejček, E., Straňák, P. a Hajič, J. (2010). Český WordNet 1.9 PDT.

Pala, K. a Ševeček, P. (1997). Valence českých sloves. In *Sborník prací FFBU*, pages 41–54, Brno.

Pala, K. a Ševeček, P. (1999). The Czech WordNet, final report. Technical report, Masarykova univerzita, Brno.

Palmer, M., Kingsbury, P. a Gildea, D. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Panevová, J. (2007). Znovu o reciprocitě. *Slovo a slovesnost*, 68(2):91–100.

Panevová, J. a Mikulová, M. (2007). On reciprocity. *The Prague Bulletin of Mathematical Linguistics*, 87:27–40.

Panevová, J. a Ševčíková, M. (2014). *Delimitation of information between grammatical rules and lexicon*, volume 215 of *Linguistik Aktuell / Linguistics Today*, pages 33–52. John Benjamins Publishing Company, Amsterdam, The Netherland.

Panevová, J. (1974). On verbal frames in Functional generative description. *The Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Panevová, J. (1980). *Formy a funkce ve stavbě české věty.* Academia, Praha.

Panevová, J. (1994). Valency frames and the meaning of the sentence. In Luelsdorff, P. A., editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Panevová, J. (1996). More remarks on control. *Prague Linguistic Circle Papers: Travaux du cercle linguistique de Prague nouvelle série*, 2:101–120.

Panevová, J., Benešová, E. a Sgall, P. (1971). *Čas a modalita v češtině.* Universita Karlova, Praha.

Popel, M. a Žabokrtský, Z. (2010). TectoMT: Modular NLP framework. In Loftsson, H., Rögnvaldsson, E. a Helgadottir, S., editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *LNCS*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L. a Weischedel, R. (2007). OntoNotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 1(4):405–419.

Ren, Z., Lü, Y., Cao, J., Liu, Q. a Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R. a Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. University of California, Berkeley.

Sag, I., Baldwin, T., Bond, F., Copestake, A. a Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania.

Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Praha.

Sgall, P., Hajičová, E. a Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/Reidel Publ. Comp., Praha/Dordrecht.

Skoumalová, H. (2001). *Czech Syntactic Lexicon*. PhD thesis, Charles University in Prague, Faculty of Arts, Prague.

Smrž, P. (2004). Quality control for WordNet development. In *Proceedings of the Second International WordNet Conference – GWC 2004*, pages 206–212, Brno, Czech Republic. Masaryk University, Brno.

Spousta, M. (2011). Featurama. http://sourceforge.net/projects/featurama/.

Spoustová, D. j. (2008). Combining statistical and rule-based approaches to morphological tagging of Czech texts. *The Prague Bulletin of Mathematical Linguistics*, 89:23–40.

Straňák, P. (2010). *Annotation of Multiword Expressions in The Prague Dependency Treebank*. PhD thesis, Univerzita Karlova v Praze, Prague, Czech Republic.

Svozilová, N., Prouzová, H. a Jirsová, A. (1997). *Slovesa pro praxi.* Academia, Praha.

Svozilová, N., Prouzová, H. a Jirsová, A. (2005). *Slovník slovesných, substantivních a adjektivních vazeb a spojení.* Academia, Praha.

Ševčíková, M. a Žabokrtský, Z. (2014). Word-formation network for czech. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. a Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tesnière, L. (1959). *Éléments de syntaxe structurale.* Klincksieck, Paris.

Tonelli, S. a Pighin, D. (2009). New features for FrameNet – WordNet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, CO, USA.

Tufiş, D., Cristea, D. a Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives. A general overview. *Romanian Journal of Information Science and Technology*, 7(1-2):9–43. Special Issue on BalkaNet.

Urešová, Z. (2011a). *Valence sloves v Pražském závislostním korpusu.* Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.

Urešová, Z. (2011b). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex).* Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.

Vadas, D. a Curran, J. R. (2007). Adding noun phrase structure to the Penn Treebank. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 240–247, Prague, Czech Republic. Association for Computational Linguistics.

Vernerová, A., Kettnerová, V. a Lopatková, M. (2014). To pay or to get paid: Enriching a valency lexicon with diatheses. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B. a Mariani, J., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2452–2459, Reykjavík, Iceland. European Language Resources Association.

Vossen, P., editor (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Norwell, MA, USA.

Žabokrtský, Z. (2005). *Valency Lexicon of Czech Verbs (PhD thesis).* PhD thesis, ÚFAL MFF UK, Praha, Czechia.