Víceslovné lexémy v syntaktickém kontextu

Alexandr Rosen — Hana Skoumalová — Jiří Znamenáček

ABSTRACT:
Multi-word lexemes in syntactic context. We start with the assumption that (i) a corpus represents the use of language, i.e. linguistic performance, (ii) a rule-based grammar represents language as a system, i.e. linguistic competence, and (iii) corpus annotation represents the interface between the two. To detect and diagnose mismatches between the language use and the language system we use a constraint-based grammar run as a constraint solver on texts tagged and dependency-parsed by stochastic tools. The texts also have MWEs (multi-word expressions) identified and transformed into a constituency-based format before the grammar is applied. We describe the role and results of the grammar, and its use to check texts annotated with morphosyntactic categories, syntactic structure and information about the status of relevant expressions as MWEs. The grammar also employs lexical resources such as a valency lexicon and a database of MWEs to make the checking more accurate and the annotation more informative. The results are represented as typed feature structures where MWE-related information can be shared by lexical and phrasal nodes. This allows for the annotation of MWEs as lexical units, independently of their analysis in terms of syntactic structure. Focusing on the interplay of MWEs with their syntactic context we analyse a number of representative examples, pointing out the pros and cons of specific solutions and the whole approach.

KLÍČOVÁ SLOVA / KEYWORDS:
čeština, HPSG, syntax, treebank, víceslovné lexikální jednotky
Czech, HPSG, syntax, treebank, multi-word expressions

1. ÚVOD

Má smysl porovnávat podobu, jakou by jazyk měl mít podle slovníku a gramatiky v idealizované podobě abstraktních hesel a pravidel, s jeho faktickým stavem podle písemných i mluvených dokladů reálného užívání? Je to možné, zajímavé a užitečné? Kladná odpověď předpokládá, že existuje metoda, která takové porovnání jazykového systému a úzu umožňuje. Cílem tohoto příspěvku je představit jeden z možných způsobů, jak takového porovnání dosáhnout, a to na příkladu víceslovných lexémů.[1]

Výběr tohoto nijak nepodstatného zákoutí jazyka není vůbec náhodný. Víceslovné lexémy lze totiž jen velmi obtížně uchopit jako lexémy nebo syntaktické konstrukce, a tedy jako součást slovníku nebo gramatiky v tradičním slova smyslu. „Uchopit" zde znamená nejen deskriptivně, teoreticky a formálně popsat, ale také softwarově implementovat v podobě konkrétního slovníku a gramatiky. K tomu přistupuje ještě

1 Pro víceslovné lexémy se často používají i jiné názvy: víceslovné lexikální jednotky, frazeologismy, kolokace, frazémy, nebo idiomy. Tyto pojmy se někdy definují s důrazem na různé specifické aspekty třídy víceslovných lexémů. Zde dáváme přednost termínu víceslovný lexém zejména z důvodů úsporného vyjádření bez nutnosti užívat zkratky.

jejich různorodá povaha a variabilita daná nejen systémovými vlastnostmi, ale také nejrůznějšími odchylkami v úzu.

Nejsme samozřejmě první, kdo se o něco takového snaží. Hlavní inspirací pro naši práci byl projekt PARSEME a bohatá literatura vzniklá na jeho základě (Sailer & Markantonatou, 2018; Markantonatou, Ramisch, Savary & Vincze, 2018; Parmentier & Waszczuk, 2019), ale již před tímto projektem se mnozí čeští i zahraniční lingvisté zabývali teoretickým popisem víceslovných lexémů (Čermák, 2007; Moon, 2007; Baldwin & Kim, 2010) i praktickým sestavováním slovníků a databází (Čermák et al., 1983–2009; Kopřivová & Hnátková, 2014), obohacováním valenčních slovníků o frazémy (Przepiórkowski, Hajič, Hajnicz & Urešová, 2017; Kettnerová, Lopatková, Bejček & Barančíková, 2018), nebo anotací víceslovných výrazů v syntakticky anotovaných korpusech (Bejček & Straňák, 2010; Lopatková, 2015).

Vycházíme z toho, že úzus je zachycen v jazykovém korpusu, zatímco systém je dán abstraktní gramatikou a slovníkem. Běžné automatické nástroje na lingvistickou anotaci textů v korpusu (taggery, parsery) dnes počítají s tím, že se jazyk v textech může více či méně odlišovat od spisovné normy, a usilují o interpretaci všech výrazů a jevů, aniž by odchylky nějak vyznačovaly. Anotace však může zachycovat úzus včetně těchto odchylek, a to i v lingvisticky definovaných pojmech. Představuje tak styčný bod mezi územ a systémem. Gramatika se slovníkem tak může text analyzovat a zjišťovat, kde a jak se úzus a systém liší.[2]

Takto představená koncepce představuje ideální stav, kdy anotace správně popisuje text a gramatika se slovníkem přesně a beze zbytku popisuje jazykový systém. Ve skutečnosti může disharmonie při analýze anotace signalizovat více možností:

1. odchylku úzu od systému, včetně odchylek, které lze považovat za chyby (tedy optimální, žádoucí výsledek);
2. chybu v anotaci textu;
3. chybu nebo opominutí v gramatice nebo slovníku.

I když stanovit diagnózu disharmonie výběrem z těchto tří možností není pro tuto metodu samozřejmé ani snadné, může být užitečný i pouhý signál, že anotace textů neodpovídá očekávání. Lze tak přispět k vývoji anotačních nástrojů, případně opravě anotace, nebo také k vývoji gramatiky a doplňování slovníku.

V tomto příspěvku se zaměříme nejprve (v části 2) na popis dat a dalších zdrojů, které využíváme v projektu Mezi slovníkem a gramatikou,[3] tedy na anotované texty, jejich formát, postup anotace a způsob porovnání s gramatikou a slovníkem. V části 3

2 Podobnou koncepci představuje spojení projektu Składnica (parsebank, tedy syntakticky anotovaný korpus — treebank, vytvořený syntaktickým analyzátorem — parserem) s projektem Świgra (gramatika) (Woliński, 2019).

3 Příspěvek vznikl v rámci prací na projektu Mezi slovníkem a gramatikou (Between Lexicon and Grammar) podpořeném Grantovou agenturou České republiky, reg. č. 16-07473S. Tento projekt navazuje na projekt Treebank češtiny na základě gramatiky, podpořený ze stejného zdroje, reg. č. 13-27184S (Jelínek et al., 2014; Jelínek, Petkevič, Rosen, Skoumalová & Vítovec, 2015; Petkevič, Rosen, Skoumalová & Vítovec, 2015; Rosen & Skoumalová, 2018).

představíme naši taxonomii víceslovných jednotek, v části 4 obsah a formu gramatiky a slovníku, v části 5 pak probereme výsledky. V závěrečné diskusi (část 6) pak zdůrazníme silné stránky i slabiny tohoto projektu.

2. DATA

Při vývoji aplikací orientovaných na přirozený jazyk je dnes běžné využívat autentická jazyková data i v lingvistickém výzkumu, zejména v podobě jazykových korpusů. Texty v nich obsažené jsou často vybaveny lingvistickou anotací, tedy např. údaji o slovním druhu a morfologických kategoriích každého slova, případně i o jeho syntaktické funkci a zapojení do syntaktické struktury věty. Někdy se anotují i frazeologismy nebo tzv. pojmenované entity, tedy slova nebo slovní spojení, která odkazují na konkrétní osoby, místa, data a události. Všechny tyto druhy lingvistické anotace se provádějí téměř vždy automaticky, a to s využitím stochastických modelů, vygenerovaných ze vzorově anotovaných textů (etalonu), případně i slovníků.

Takové metody mohou být dostatečně spolehlivé pro řadu účelů, ale jen výjimečně jsou zcela bez chyb. Navíc jsou postaveny před úkol anotovat (tj. analyzovat) reálné texty, které někdy neodpovídají nejen jazykové normě, ale ani etalonu. I s takovými případy se však tyto metody vypořádají, aniž by na výsledku bylo na první pohled patrné, že text neodpovídá očekávání. To může být často užitečné, pokud výsledek využívají prakticky zaměřené aplikace, ale ne vždy žádoucí, pokud nás zajímají právě odchylky od libovolně definovaného standardu.

Důraz na empirická data a možnosti jejich využití pro praxi i výzkum se tak střetává s teoreticky motivovanou potřebou verifikovat hypotézy formálně vyjádřené pravidly gramatiky a slovníkem. Nespočetné doklady jazykového úzu se sice stávají součástí empirických modelů jazyka, ale tyto modely nejsou z hlediska lingvistické teorie dostatečně transparentní, nepracují s lingvistickými pojmy a nelze v nich ověřovat lingvistické hypotézy. Výsledkem aplikace empirických modelů na texty však je anotace, kterou lze ověřovat a analyzovat pomocí gramatiky a slovníku vytvořených tradičními lingvistickými metodami. Anotace korpusu tak představuje styčný bod mezi langue a parole, mezi jazykem jako systémem v podobě gramatiky a slovníku na jedné straně a jeho užitím, který představuje korpus, na straně druhé (Rosen, 2018).

Víceslovné lexémy definujeme široce jako výrazy, které se skládají z více slovních tvarů a vyznačují se alespoň některým z více typů idiomatičnosti. Z hlediska jazykového systému se nacházejí mezi dvěma póly: slovníkem a gramatikou. Prostor mezi těmito póly je zaplněn víceslovnými lexémy nejrůznějších typů, které se jim navíc blíží či vzdalují v závislosti na úhlu pohledu. Ustrnulá spojení jako např. *třesky plesky* se výrazně odlišují od konstrukcí typu *honit vodu*, a to z důvodů morfologických, syntaktických, sémantických i pragmatických. Když se omezíme jen na formálně snadno pozorovatelné projevy těchto odlišností, tedy např. míru oné „ustrnulosti", snadno se ujistíme, že kontext nemá na podobu ustrnulých spojení žádný vliv. Oproti tomu víceslovné lexémy, které lze samy o sobě analyzovat jako syntaktické konstrukce s více či méně pravidelnými vlastnostmi svých členů, se často začleňují do kontextu podobně jako syntaktické konstrukce formálně identické. Příklad (1) ukazuje triviální

možnost flexe a negace slovesa ve víceslovném lexému. Homonymní doslovné užití týchž lexémů v příkladu (2) vykazuje stejné formální vlastnosti. V příkladu (3) se spojuje idiomatické a doslovné čtení v jedné koordinaci.

(1) *Tihle kluci* ***nehoní vodu*** *v naleštěných meďourech.*[4]
(2) *Čeřidla* ***honí vodu*** *kolem lodí, aby nezamrzla.*[5]
(3) *Praktická Lundová ale* ***nehoní vodu****, nýbrž vraha. Její ohoz je stále stejný: svetr, džíny, holiny.*[6]

Příklad (4) ukazuje, že víceslovné lexémy připouštějí vnitřní modifikaci, která může zesilovat nebo jinak kvalifikovat význam celého slovního spojení. Tento příklad lze interpretovat i jako kombinaci spojení *honit vodu* a *velká voda*. Příklad (5) ukazuje slovoslednou variaci a příklad (6) možnost diateze řídícího slovesa.

(4) *Ve mně by Sally potkala kluka, který přijel do Prahy z Plzně* ***honit velkou vodu****, což se mu moc nepovedlo.*[7]
(5) *Rychtecký by na lázeňské kolonádě v tomhle oblečku* ***vodu honit*** *nemohl.*[8]
(6) *V Praze se také* ***honí voda****;* […][9]

Všechny výše uvedené příklady dokládají značnou variabilitu některých víceslovných lexémů. Tato variabilita je přitom alespoň u některých víceslovných jednotek srovnatelná s jejich neidiomatickými obdobami, a jde tedy o variabilitu očekávanou, která je dána propojením lexikálních a gramatických specifikací. Proto je důležité modelovat víceslovné lexémy v teoretickém a formálním rámci, který spojení slovníku a gramatiky umožňuje.

Kromě této „očekávané" variability, dané postavením víceslovných lexémů v systému jazyka, existuje i variabilita daná užíváním jazyka, kterou víceslovné lexémy sdílejí se všemi výrazy a jevy, jako jsou třeba „chyby" způsobené performančními faktory v užívání jazyka nebo tvůrčí modifikace standardu. Podobně jako u jiných výrazů a konstrukcí lze proti sobě postavit systém a úzus a konfrontovat anotaci korpusu s gramatikou a slovníkem.

3. TAXONOMIE VÍCESLOVNÝCH VÝRAZŮ

Abychom mohli víceslovné jednotky popsat a rozpoznat v různých podobách a kontextech, je nutné je klasifikovat podle více hledisek:

4 SYNv8 (Křen et al., 2019); *Aha!*, 21. 10. 2010.
5 SYNv8; *Mladá fronta DNES*, 2. 3. 2018, Liberecký kraj.
6 SYNv8; *Magazín Víkend DNES*, č. 13/2013.
7 SYNv8; *Vlasta*, č. 22/2011.
8 SYNv8; *Bulvár*, č. 3/2014.
9 Karel Čapek: Ze života slov, *Lidové noviny*, 9.1.1934; http://ld.johanesville.net/capek-80-o-umeni-a-kulture-iii?page=236.

1. typ užití (např. přísloví, pranostika, přirovnání, citace, termín, víceslovné synsémantikum)
2. syntaktický typ (např. fráze jmenná, adjektivní, adverbiální, předložková, slovesná plnovýznamová nebo s kategoriálním slovesem, složená spojka nebo předložka)
3. míra ustrnulosti (možnost užití variant nebo fragmentů základní podoby, slovosledná a morfologická omezení a omezení slovesných transformací)
4. typ idiomaticity (lexikální, morfologická, syntaktická, sémantická, pragmatická, statistická).[10]

Kromě toho určujeme u víceslovných lexémů také jejich lemma, definici, zařazení podle stylu nebo registru jazyka, syntaktickou strukturu (strukturní vzorec, závislostní a složkový strom) a valenci. Až na valenci uvádíme jen specifikace, které jsou z hlediska standardní gramatiky neočekávané, takže např. u spojení *Pandořina skříňka* není uvedeno, že se nevyskytuje v obráceném slovosledu, protože to platí obecně pro všechny jmenné fráze podobného typu, tedy např. i pro *dědečkovy brýle*. Všechny tyto údaje jsou uvedeny v lexikální databázi, s níž mohou pracovat uživatelé i aplikace. Podrobnější popis databáze uvádí Hnátková et al. (2019).

4. POSTUP ANOTACE

Databáze je jedním z více zdrojů, které se při anotaci využívají. Další z nich jsou valenční slovník (Lopatková, Kettnerová, Bejček, Vernerová & Žabokrtský, 2016; Rosen & Skoumalová, 2018), tagger (Skoumalová, Hnátková & Petkevič, 2011) a závislostní parser (Jelínek, 2016), nástroj pro automatickou anotaci víceslovných lexémů v textu (Hnátková, 2002) a v neposlední řadě sada nástrojů, které tvoří páteř celého postupu automatické anotace a zajišťují konverzi mezi různými formáty anotace a textových dat (Jelínek et al., 2014). Text se anotuje tímto postupem (některé méně podstatné kroky neuvádíme):

1. morfologická analýza (tagger);
2. anotace víceslovných lexémů;
3. závislostní syntaktická analýza (parser);
4. doplnění údajů o víceslovných lexémech z databáze;
5. konverze do formátu typovaných sestav rysů a složkových stromů;

10 Pod idiomaticitou rozumíme odlišnost od standardního systémového chování celého víceslovného lexému nebo jednotlivých slov v něm obsažených. Jednotlivé typy idiomaticity se mohou vzájemně kombinovat, s výjimkou idiomaticity statistické dané obligatorností a proximitou, u které má víceslovný lexém vždy kompozicionální význam, a idiomaticity sémantické, kde možnost doslovného (kompozicionálního) čtení vylučujeme. Mezi víceslovné lexémy zahrnujeme pouze ty n-tice slov, které tvoří frázi, anebo složené synsemantikum. Neuvažujeme tedy taková frekventovaná spojení jako *by se* nebo *bychom zítra*. (Více viz Cvrček, 2013; Hnátková et al., 2017; Hnátková et al., 2018.)

6. doplnění údajů z valenčního slovníku;
7. ověření a doplnění anotace pomocí gramatiky.

Při tomto postupu na sebe navazují samostatné moduly, které mají za úkol provést vždy jeden krok anotace. Tím se náš přístup liší od jiných projektů, ve kterých byla identifikace víceslovných jednotek součástí gramatiky, ať už se jednalo o systémy založené na HPSG (Sailer & Richter, 2002; Richter & Sailer, 2014), nebo na LFG (Dyvik, Losnegaard & Rosén, 2019).

V následující části přiblížíme gramatiku, její funkci, obsah a formu.

5. GRAMATIKA

Úkolem gramatiky v tomto projektu je v první řadě ověřit, zda úzus, zastoupený lingvisticky anotovanými texty s důrazem na víceslovné lexémy, odpovídá pravidlům této gramatiky a lexikálním specifikacím ze slovníku, společně představujícím jazyk jako systém. Negativní výsledek, tedy zjištění, že anotace textů neodpovídá pravidlům, může v konkrétním případě znamenat i nesprávnou anotaci nebo neadekvátní popis v gramatice či slovníku.

Anotace, která obsahuje morfologickou a syntaktickou analýzu textů spolu s identifikací víceslovných lexémů, prochází kontrolou formální i jazykové správnosti a konzistence. Anotace vyjadřující kompatibilní analýzu jsou obohaceny o další informace: relevantní vlastnosti lexikálních kategorií jsou promítnuty do frázových uzlů, lexikální kategorie dostávají valenční rámce, které mají být naplněny větnými členy v dané větě. U víceslovných lexémů, které jsou jako takové anotovány jen v terminálních uzlech, se relevantní informace o daném lexému dostávají do vyšších složek. Toto sdílení informací o víceslovném lexému mezi syntaktickými dcerami a matkami končí u nejnižší složky, která obsahuje všechny součásti frazému, i když může obsahovat i jiné podsložky.

Kromě syntaktického modulu, který kontroluje syntaktickou strukturu a morfosyntaktické údaje obsažené v anotaci textů a tuto analýzu doplňuje o informace z valenčního slovníku a z terminálních uzlů složkového stromu, je součástí gramatiky také modul lexikální, který využívá lexikální hesla z externího valenčního slovníku a generuje valenční rámce pro slovesa v textu, a to ve více verzích podle možných diatezí (viz Skoumalová, 2016).

Gramatika je implementována v systému Trale,[11] formalismu určeném pro gramatiky založené na HPSG, lingvistické teorii pro modelování lingvistických výrazů jako strukturovaných sestav rysů (viz např. Pollard & Sag, 1994). Formát gramatického formalismu je kompatibilní s formátem analýzy a slovníkových hesel. Jednou z nejdůležitějších vlastností této teorie i formalismu je pojetí gramatiky jako množiny omezení, která jsou kladena na výrazy jazyka. Nejde tedy o množinu pravidel, která v procedurálním smyslu pomocí syntaktických stromů generují nebo analyzují ře-

11 https://hpsg.hu-berlin.de/Software/Trale/, http://www.sfs.uni-tuebingen.de/hpsg/archive/projects/trale

tězce slov. Výhodou takto pojaté gramatiky je kromě jiného i možnost ji interpretovat zároveň jako návod k syntéze (generování) i analýze. V našem projektu ji však interpretujeme právě jen jako omezení kladená na morfologickou a syntaktickou anotaci textu, včetně anotace víceslovných lexémů. Tím se nejvíc liší od typických aplikací lingvisticky formulovaných gramatik. Gramatika dostává na vstupu sestavy rysů, vytvořené stochastickými nástroji pro morfologickou a syntaktickou analýzu, doplněné o specifikace víceslovných lexémů a valenčních rámců a zkonvertované do náležité podoby. Proto neobsahuje žádná syntaktická pravidla bezkontextového typu a funguje jako constraint solver,[12] přičemž omezení pocházejí ze tří zdrojů: data, slovník a vlastní gramatika.

Každá věta je na vstupu anotována jednoznačně, tj. odpovídá jí právě jeden syntaktický strom, který ani v jednom uzlu neobsahuje žádné nejednoznačné údaje. Víceznačné interpretace mohou vzniknout jen v důsledku podrobnější taxonomie morfosyntaktických kategorií nebo nejistoty ohledně výběru valenčního rámce.

6. ANALYZOVANÉ PŘÍKLADY

Na jedenácti větách, z nichž většina obsahuje alespoň dva víceslovné lexémy, ukážeme výsledky aplikace gramatiky na anotované texty a problémy s tím spojené. Přitom se zaměříme hlavně na analýzu víceslovných lexémů v syntaktickém kontextu. U každé věty uvedeme její složkovou strukturu na vstupu do gramatiky, někdy ukážeme i údaje u jednotlivých relevantních uzlů po průchodu gramatikou. Je důležité si uvědomit, že syntaktická struktura je dána stochastickou závislostní analýzou, transformovanou do složkového tvaru. Gramatika ji nemůže měnit, může ji jen kontrolovat a doplňovat údaje do uzlů struktury.

Příklady obsahují víceslovné lexémy několika syntaktických typů. Uvádíme jejich seznam spolu s čísly příkladů, označením případné sémantické idiomatičnosti pomocí hvězdičky a doplněním typu příslovečného nebo přívlastkového určení (za znaménkem plus).

Jmenná fráze
- *široká veřejnost* (7)
- *sametová revoluce** (9)
- *celou dobu* (10) +čas
- *děs a hrůza* (11)
- *klinická smrt* (17) +termín
- *svěcená voda* (17)
- *neřád neřádská* (18)

12 Constraint solver je algoritmus a/nebo implementace algoritmu, který hledá řešení splňující všechna daná omezení. U gramatiky to tedy neznamená, že gramatika něco analyzuje nebo generuje, ale aplikuje omezení daná gramatickými pravidly a slovníkem na vstupní data. Výsledkem je jednak verdikt o souladu vstupních dat s předpoklady gramatiky a slovníku, jednak potenciálně informativnější obsah výstupních dat.

Předložková fráze
v dnešní době (7) +čas
při vší úctě (9) +okolnost
před koncem (12) +čas
*na plné obrátky** (12) +způsob
před nedávnem (13) +čas
na míru (14) +způsob
v řadě případech (19) +okolnost

Slovesná fráze s kategoriálním slovesem
mít nárok (8)
dávat najevo (10)
podat výkon (13)
přivodit smrt (13)
*být nesvůj** (14)

Slovesná fráze plnovýznamová
chodit do práce (8)
*lhát si do kapsy** (9)

Složená spojka
i když (8) (11)

Složená předložka
*v duchu** (9)

Adverbiální fráze
pomalu, ale jistě (9) +způsob

Následuje rozbor jednotlivých příkladů.

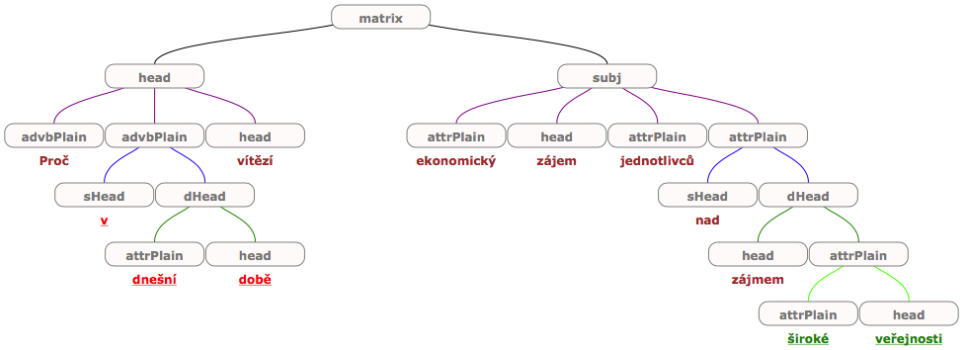
(7) *Proč **v dnešní době** vítězí ekonomický zájem jednotlivců nad zájmem **široké veřejnosti**?*[13]

Příklad (7) obsahuje dva víceslovné lexémy, zvýrazněné v Grafu 1 podtržením. Každý z nich má v syntaktické analýze svoji složku, přičemž první z nich obsahuje další neterminální uzel pro jmennou frázi. Na vstupu do gramatiky jsou však údaje o víceslovných lexémech uvedeny jen u terminálních uzlů. Až gramatika je promítne do všech vyšších složek.[14]

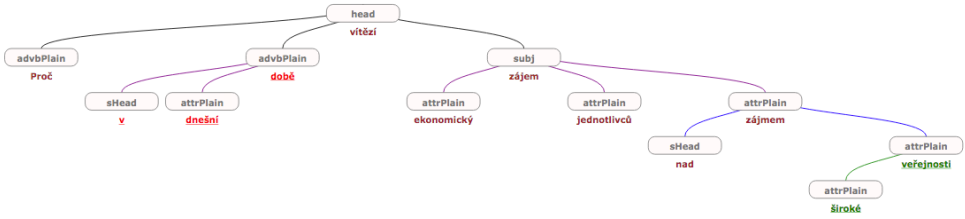
Z Grafu 1 je už při zběžném prozkoumání patrné, že do subjektové části věty je chybně zahrnuta i předložková fráze *nad zájmem veřejnosti*. Jde o chybu stochastického parseru, na niž by gramatika jako na možný problém mohla upozornit na zá-

13 SYNv8; *Deníky Bohemia*, 8. 9. 2009.

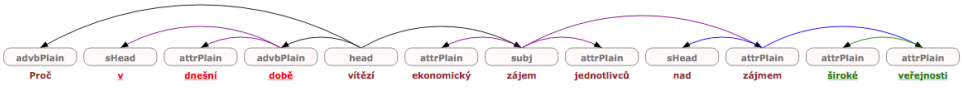
14 Syntaktická struktura v Grafu 1 ani v dalších grafech neobsahuje interpunkci.



GRAF 1: Složková struktura věty (7).



GRAF 2: Závislostní struktura věty (7) s hloubkovými závislostmi vygenerovaná ze složkové struktury.

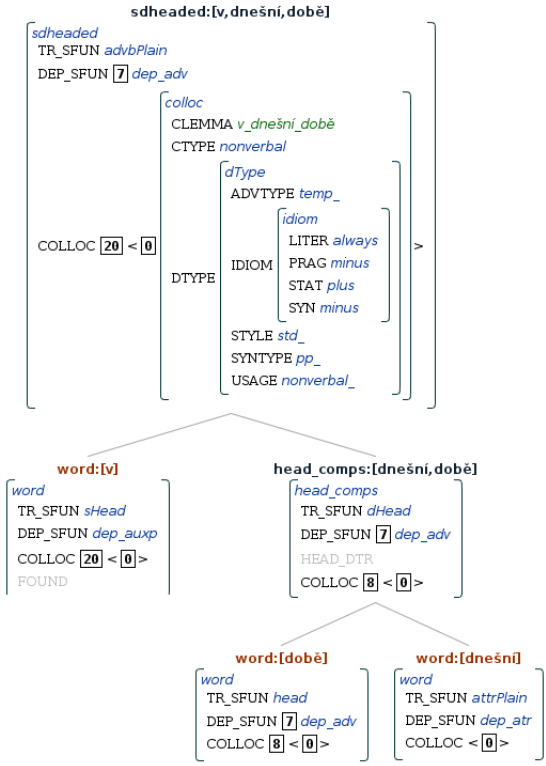


GRAF 3: Závislostní struktura věty (7) s povrchovými závislostmi v linearizované podobě.

kladě valenčního rámce slovesa *vítězit*, vybaveného slotem pro předložkovou frázi s předložkou *nad*.[15]

Uzly stromu jsou označeny nikoli frázovou kategorií (NP, PP), ale funkcí: matrix je funkce pro nejvyšší uzel, reprezentující celou větu, head pro hlavu, tj. řídící větný člen (ve větě pro slovesnou frázi nebo sloveso, ve jmenné frázi pro řídící substantivum), subj pro podmět, advbPlain pro adverbiale (příslovečné určení), attrPlain pro atribut (přívlastek). Analytické tvary se synsémantiky (pomocnými slovy) se dělí na sHead (povrchovou hlavu), představující synsémantickou část, a dHead (hloubkovou hlavu), zastupující autosématickou část tvaru. Díky tomuto řešení lze tutéž analýzu ze stejných dat zobrazit také jako závislostní strom, obsahující pouze terminální uzly. Je možné si vybrat mezi závislostmi hloubkovými (Graf 2), kde synsémantika jsou závislá na hloubkové hlavě, nebo povrchovými (Graf 3), kde je tomu naopak, a různými

15 Problém analýzy předložkových frází v podobném kontextu se objevuje i dále v příkladu (9).



GRAF 4: Složková analýza spojení *v dnešní době*.

způsoby jejich zobrazení: strom bez důsledného ohledu na slovosled (Graf 1), strom s uzly slovosledně uspořádanými (Graf 2) a strom linearizovaný (Graf 3).

Údaje o víceslovných lexémech jsou obsaženy v příslušných uzlech stromových struktur, které obsahují řadu dalších údajů ve tvaru typovaných sestav rysů. Graf 4 ukazuje vybrané údaje na složkové analýze spojení *v dnešní době*.[16] Každý uzel stromu je zde označen typem složky (sdheaded — fráze složená z povrchové a hloubkové hlavy, head_comps — fráze složená z řídící hlavy a komplementů, word — lexikální položka, tj. terminální uzel) a výrazem, který reprezentuje. Údaje pro každý uzel jsou pak vyjádřeny sestavou rysů, tj. množinou dvojic ATRIBUT *hodnota*, přičemž *hodnota* může být další vnořená sestava rysů. V grafu jsou kromě údajů o víceslovném lexému (COLLOC) uvedeny syntaktické funkce z původní závislostní analýzy (DEP_SFUN) a po konverzi do složkové struktury (TR_SFUN). CLEMMA je lemma víceslovného lexému (obvykle vycházející z jeho základního tvaru), CTYPE je jeho typ podle třídění použitého např. v korpusu SYNv8.[17] Údaje uvedené v DTYPE jsou

16 Grafické rozhraní Grale, které zobrazuje podrobné analýzy provedené gramatikou, umožňuje ukázat nebo skrýt libovolné části struktur.

17 https://wiki.korpus.cz/doku.php/cnk:syn:verze8, https://wiki.korpus.cz/doku.php/seznamy:frazemy, Křen et al. (2019)

převzaté z databáze jako kategorie užitečné pro anotaci: ADVTYPE udává typ příslovečného určení (pokud lexém lze takto interpretovat), STYLE stylové zařazení (zde spisovná čeština), SYNTYPE syntaktickou kategorii (zde předložková fráze), USAGE způsob užití, tj. tradiční klasifikace (zde „neslovesný frazém"). Hodnota atributu IDIOM specifikuje typy idiomatičnosti, tj. v jakém smyslu je víceslovný lexém idiomatický.[18] Hodnota atributu LITER uvádí, jak často lze lexém interpretovat doslovně. Je to vlastně obrácená míra sémantické idiomatičnosti. U výrazu *v dnešní době* jde tedy vždy o doslovné, kompozicionální vyjádření, výraz tedy není nikdy sémanticky idiomatický. Není idiomatický ani pragmaticky (PRAG, na rozdíl třeba od pozdravu *dobrý den*), ani syntakticky (SYN, na rozdíl třeba od *padni komu padni*), je však idiomatický statisticky, tj. toto spojení se vyskytuje výrazně častěji než např. *v dnešním čase*. Hodnota atributu COLLOC je seznam, který může obsahovat údaje i o více než jednom víceslovném lexému, což se hodí v případě, že část jednoho lexému je zároveň součástí jiného. To se může stát, když se víceslovné výrazy překrývají nebo jeden obsahuje druhý. Sdílení údajů o tom, že uzel reprezentuje daný lexém nebo jeho část, je zajištěno koindexací čísly v rámečcích. Tak např. údaje o víceslovném lexému *v dnešní době* uvedené u nejvyššího uzlu jsou díky koindexaci pomocí indexu 0 dostupné ve všech ostatních uzlech stromu.

Další příklad (8) je trochu komplikovanější: obsahuje tři víceslovné lexémy, z nichž pouze jeden je ve struktuře analyzován jako jedna složka: *nechodí do práce*.

(8) ***I když* nechodí do práce**, ***mají*** *totiž* ***nárok*** *na příspěvek na bydlení.*[19]

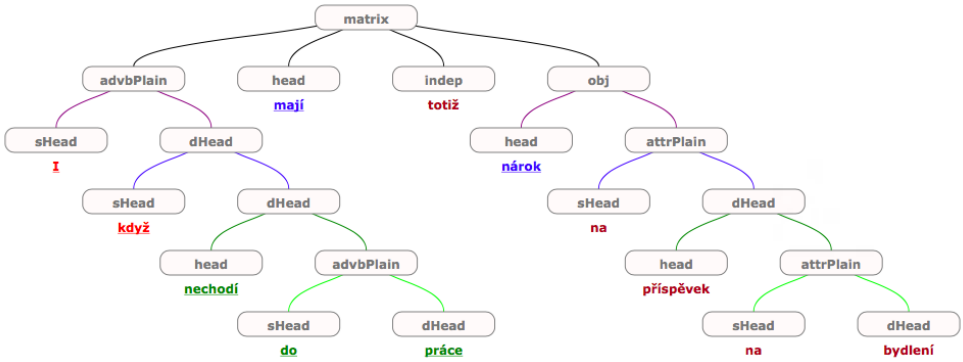
Výraz *mají nárok* je ve větě slovosledně rozdělen slovem *totiž* a strukturně modifikací substantiva *nárok* předložkovou frází *na příspěvek na bydlení* (Graf 5). Jde o syntakticky korektní analýzu. Všechny části lexému *mít nárok* jsou však obsaženy až v nejvyšší složce, reprezentující celou větu — matrix. Součástí anotace uzlu matrix jsou také údaje o tomto lexému, včetně odkazů na všechny jeho části. Anotace víceslovných lexémů je tedy nezávislá na syntaktické struktuře. To je vidět i na třetím víceslovném lexému v této větě, složené spojce *i když*. Terminální uzly pro obě části spojení jsou jako povrchové hlavy (sHead) součástí dvou různých složek, což pro *i když* jako víceslovný lexém typu složená spojka není optimální analýza. Na tento fakt upozorňuje údaj o typu víceslovného lexému ve složce advbPlain obsahující obě části složené spojky (Graf 5).[20]

Následující příklad (9) obsahuje hned pět víceslovných lemmat: okolnostní určení vyjádřené předložkovou frází *při vší úctě*, určení způsobu vyjádřené koordinací adverbií *pomalu, ale jistě*, označení historické události vyjádřené jmennou frází *sametová*

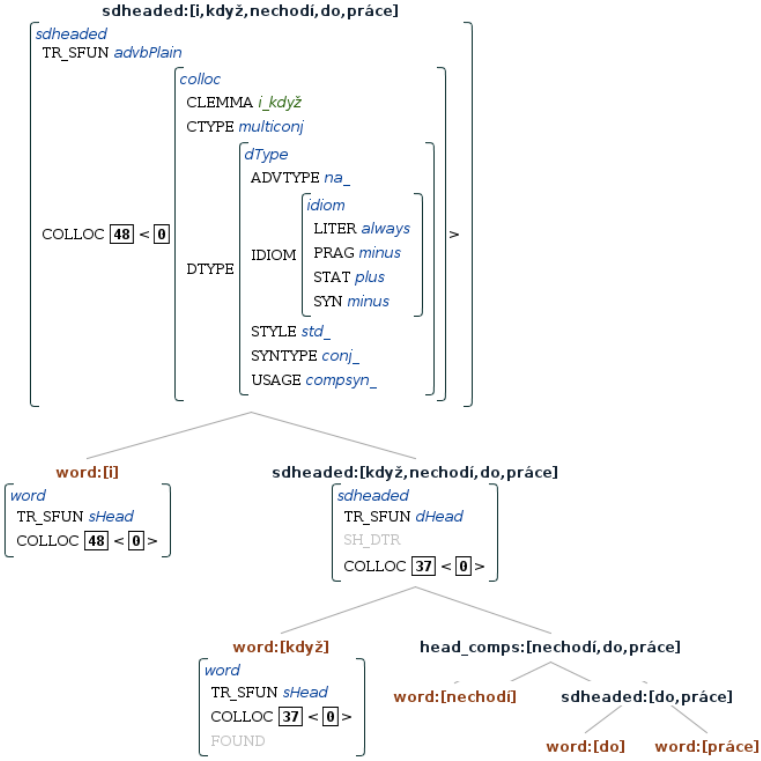
18 Z důvodů úspory místa není v grafu uvedena idiomatičnost lexikální a morfologická, kterými se v tomto článku nezabýváme. V databázi víceslovných lexémů jsou však samozřejmě uvedeny a gramatika s nimi počítá.

19 SYNv8; *Lidové noviny*, 11. 10. 2014.

20 USAGE *compsyn* jako charakteristika složené spojky *i když* v Grafu 6 označuje toto spojení jako víceslovné synsémantikum, CTYPE *multiconj* a SYNTYPE *conj* jej označují jako složenou spojku.

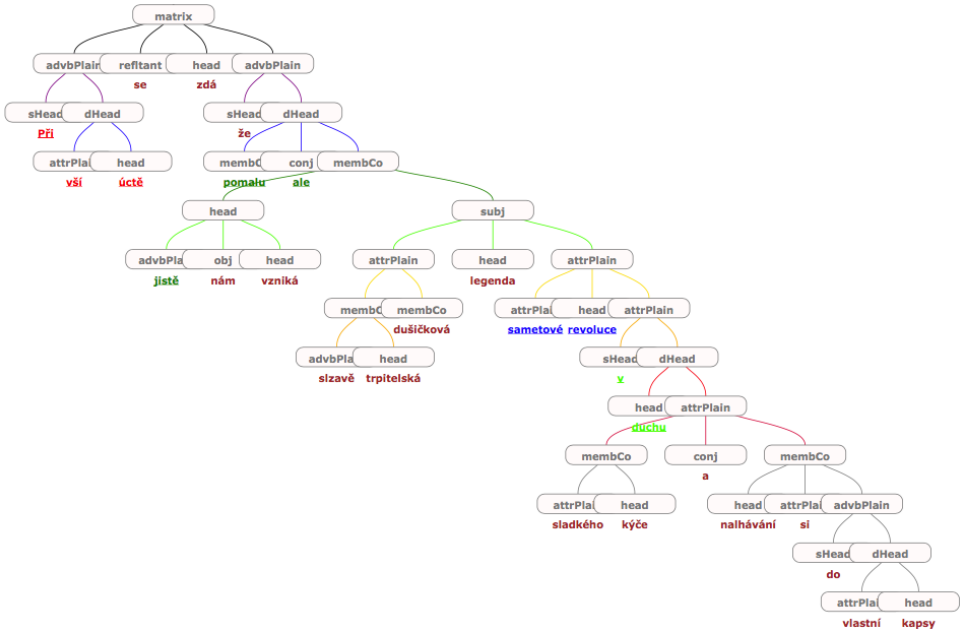


GRAF 5: Složková struktura věty (8).



GRAF 6: Složková analýza složené spojky *i když*.

revoluce, složenou předložku *v duchu* a spojení *nalhávání si do vlastní kapsy*, odvozené od sémanticky idiomatické slovesné fráze *lhát si do kapsy* v nominalizované podobě lexikální varianty slovesa (*lhát si* → *nalhávat si* → *nalhávání si*) s přidanou modifikací (*vlastní*).



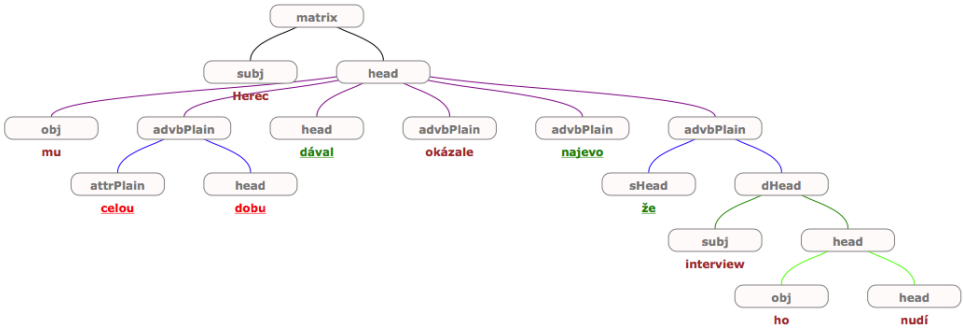
GRAF 7: Složková struktura věty (9).

(9) ***Při vší úctě*** *se zdá, že* ***pomalu, ale jistě*** *nám vzniká slzavě trpitelská, dušičková legenda* ***sametové revoluce*** **v duchu** *sladkého kýče a* ***nalhávání si do vlastní kapsy***.[21]

Spojení *při vší úctě* a *nalhávání si do vlastní kapsy* jsou celá obsažena každá ve své složce, bez dalšího materiálu. Výraz *sametová revoluce* je (ve stejné složce) modifikován rozvitou předložkovou frází, což je z hlediska syntaktické struktury i charakteristiky víceslovného spojení správná analýza.[22] Složená předložka *v duchu* je v tomto příkladu analyzována podobně jako složená spojka *i když* z příkladu (8), která byla ve složkovém stromě rozdělena do dvou složek proti své povaze jako složené synsémantikum, ale ne zcela proti duchu složkové analýzy. Zde lze dokonce argumentovat, že dvojí pohled na spojení *v duchu* je oprávněný a v analýze korektně reprezentovaný. Lexikálně motivovaná charakteristika spojení jako složené předložky je vyjádřena hodnotou atributu COLLOC v příslušných uzlech stromu, včetně uzlů neterminálních. Současně je syntakticky motivovaná analýza, podle níž je substantivum *duch* modifikováno koordinací genitivních jmenných frází, vyjádřena syntaktickými funkcemi

21 SYNv8; *Reflex*, č. 47/2004.

22 I když je možné pracovat i s variantou, že předložková fráze *v duchu…* modifikuje sloveso, respektive slovesnou frázi *vzniká…* . Jde o známý problém s analýzou předložkových frází v postavení za slovesem a substantivem (PP-attachment).



GRAF 8: Složková struktura věty (10).

a strukturně.[23] Věta v příkladu (9) však obsahuje také spojení *pomalu, ale jistě*, které je analyzováno syntakticky jako koordinace adverbia *pomalu* s celým zbytkem věty (Graf 7). Tato analýza je zcela chybná z každého pohledu, ale díky identifikaci tohoto víceslovného spojení a označení jeho typu v příslušných uzlech stromu lze tuto chybu syntaktické analýzy detekovat a případně i opravit.

Příklad (10) obsahuje dva víceslovné lexémy: jmennou frázi ve funkci časového určení *celou dobu* a slovesnou frázi s kategoriálním slovesem *dávat najevo*.

(10) *Herec mu* ***celou dobu dával*** *okázale* **najevo**, *že ho interview nudí.*[24]

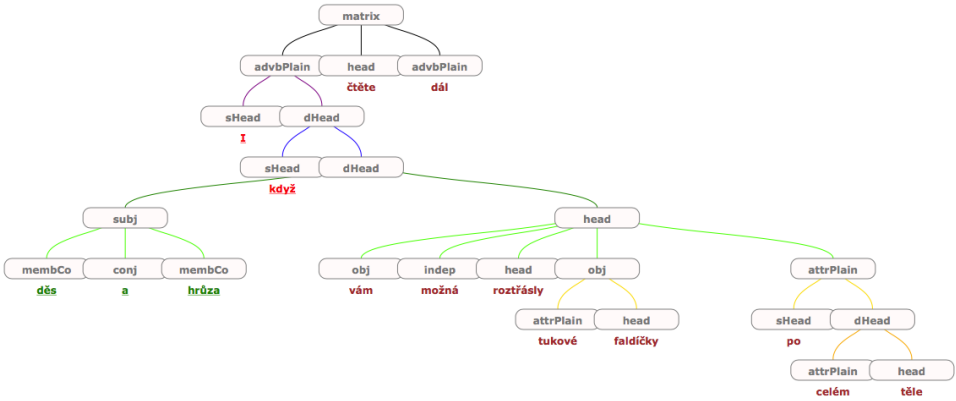
U spojení *celou dobu* se syntaktická analýza shoduje s anotací víceslovných lemmat: *celou dobu* je v obou případech označeno jako příslovečné určení a je celé samo v jedné složce (Graf 8). Spojení *dávat najevo* je syntakticky analyzováno bez ohledu na jeho verbonominální povahu jako adverbiální modifikace slovesa.[25] Jde tedy o analýzu analogickou s výše uvedenými příklady složené spojky *i když* a předložky *v duchu*: části víceslovného lexému nejsou v syntaktické struktuře jedinými součástmi jedné složky, ale jsou strukturně i funkčně začleněny do struktury na základě svých vlastností jako samostatných lexémů. Jak již bylo uvedeno u předchozího příkladu, vzhledem k identifikaci a kategorizaci všech součástí víceslovných lexémů i v relevantních vyšších složkách lze považovat toto řešení za adekvátní vyjádření lexikálního a syntaktického pohledu na tyto typy víceslovných lexémů.

Spojení *dávat najevo* je zde však zajímavé ještě z jednoho důvodu. Pokud by syntaktická analýza ignorovala jeho verbonominální status a brala v úvahu valenční vlastnosti slovesa *dávat*, strom v Grafu 8 by nebylo možné sestavit. Doplnění slovesa *dávat* vedlejší větou uvozenou spojkou *že* není možné bez lexému *najevo*. Stochastická

23 Databáze víceslovných lexémů obsahuje také spojení *v duchu* jako určení způsobu vyjádřené předložkovou frází: *v duchu si říkala, že musí zachovat klid*. Volba mezi těmito dvěma možnostmi je v daném kontextu usnadněna jejich odlišnými syntaktickými vlastnostmi.

24 SYNv8; *Rytmus života*, č. 30/2013.

25 Termín „verbonominální" používáme z praktických důvodů přesto, že přísně vzato nejde o spojení slovesa a jména, takže jeho použití není zcela na místě.



GRAF 9: Složková struktura věty (11).

syntaktická analýza, jejímž výsledkem je závislostní struktura, která se po transformaci na strukturu složkovou kontroluje a doplňuje gramatikou, s valencí explicitně nepracuje. Nesoulad mezi lexikální specifikací valence ve valenčním slovníku a syntaktickou strukturou se tak projeví až při kontrole gramatikou. Pokud pokus uplatnit některý z valenčních rámců selže, gramatika tuto skutečnost oznámí a ukáže částečně zkontrolovanou a doplněnou strukturu, což se v tomto příkladu také stalo. Valenční slovník totiž zatím neobsahuje specifikace rámců pro víceslovné lexémy, i když databáze víceslovných lexémů valenční rámce obsahuje, např. u *dát najevo* je uvedeno povinné doplnění větnými členy vyjadřujícími nějakou faktickou skutečnost, včetně vedlejší věty uvozené spojkou *že*.

Věta (11) spolu s Grafem 9 pak ukazuje kromě již výše zmíněné analýzy složené spojky *i když* také bezproblémové substantivní spojení *děs a hrůza*.

(11) ***I když*** *vám možná* ***děs a hrůza*** *roztřásly tukové faldíčky po celém těle, čtěte dál.*[26]

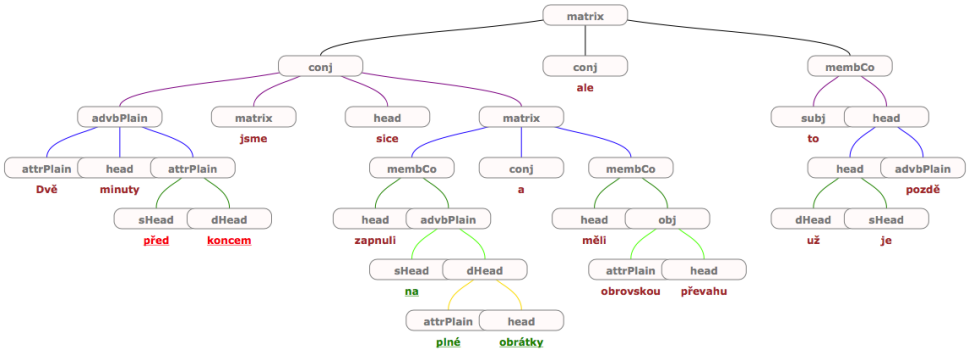
Příklad (12) ukazuje dva víceslovné lexémy vyjádřené předložkovými frázemi: časové určení *před koncem* ve funkci přívlastku a určení způsobu *na plné obrátky* v příslovečné funkci. Spojení *na plné obrátky* je zde užito v sémanticky idiomatickém smyslu, i když tomu tak nemusí být vždycky. V anotaci zatím způsob užití takto homonymních víceslovných lexémů nerozlišujeme.

(12) *Dvě minuty* ***před koncem*** *jsme sice zapnuli* ***na plné obrátky****, a měli obrovskou převahu, ale to už je pozdě.*[27]

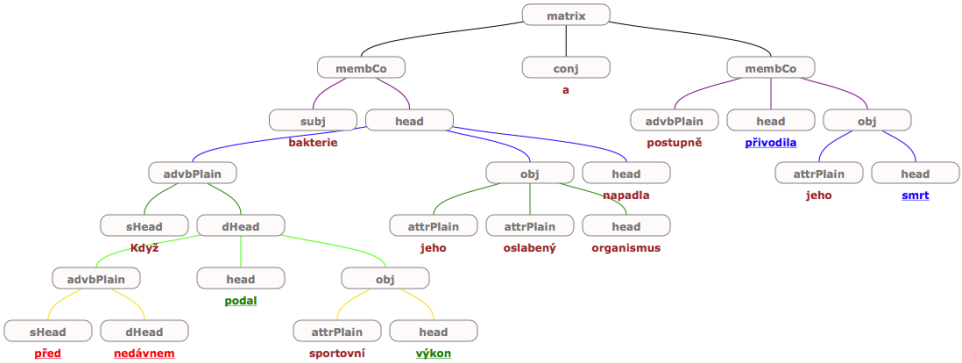
V příkladu 13 jsou kromě časového určení vyjádřeného předložkovou frází *před nedávnem* také dvě slovesné fráze s kategoriálním slovesem *podat výkon* a *přivodit smrt*, v obou případech s modifikací substantiva.

26 SYNv8; *Blesk*, 6. 12. 1999.

27 SYNv8; *Deníky Moravia*, 12. 4. 2014.



GRAF 10: Složková struktura věty (12).



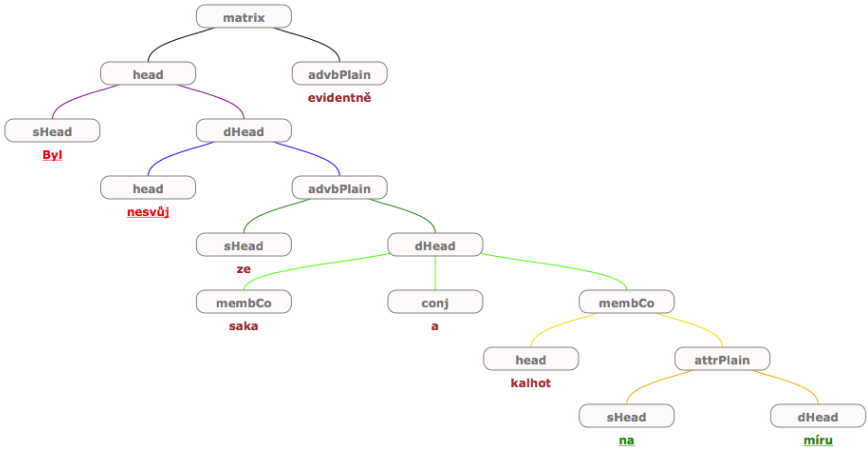
GRAF 11: Složková struktura věty (13).

(13) *Když **před nedávnem podal** sportovní **výkon**, bakterie jeho oslabený organismus napadla a postupně **přivodila** jeho **smrt**.*[28]

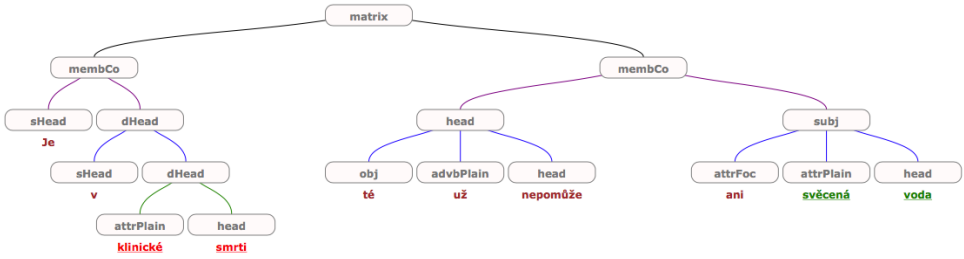
Alternativní syntaktická analýza se slovesem a substantivem v jediné, o další větné členy nerozšiřitelné složce je v těchto případech těžko představitelná (Graf 11). Považujeme to za další argument ve prospěch stávajícího řešení s lexikálním a syntaktickým pohledem na víceslovné lexémy.

Příklad (14) obsahuje jedno z více možných spojení se slovem *nesvůj*, které se kromě *být* může pojit také se slovesy *cítit se*, *připadat si* nebo *vypadat*, případně být použito samostatně ve stejném významu (15, 16). S variabilitou součástí lexému se v databázi počítá, ale možnost užít slovo samostatně naráží na definici víceslovných lexémů. Kromě možnosti postulovat v těchto případech elipsu přechodníku v syntaktické struktuře se jako řešení nabízí možnost připustit jako variantu víceslovného lexému i jediné slovo. Tuto druhou možnost v rámci daného formalismu preferujeme.

28 SYNv8; *Právo*, 3. 12. 2002.



GRAF 12: Složková struktura věty (14).



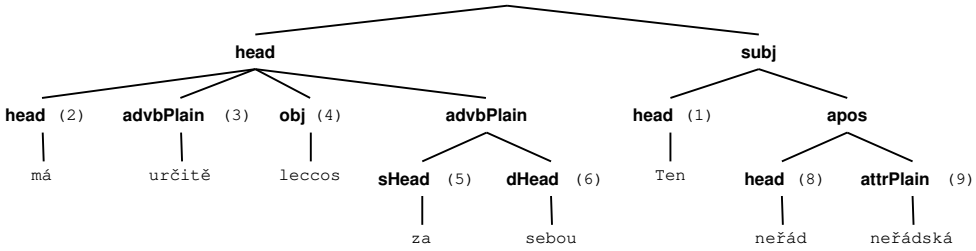
GRAF 13: Složková struktura věty (17).

(14) ***Byl*** *evidentně* ***nesvůj*** *ze saka a kalhot* ***na míru***.[29]
(15) *Je tam ticho, hostů je málo a personál stojí podél zdí,* ***nesvůj*** *z nedostatku práce.*[30]
(16) *Trochu* ***nesvůj*** *si uvědomil, že se tam s ním možná bude zacházet o něco přísněji než v tomto světě.*[31]

Spojení *na míru* jako příslovečné nebo přívlastkové určení způsobu může být užito také jako složená předložka a platí pro něj analogicky to, co bylo řečeno o spojení *v duchu* výše v příkladu (9).

Příklad (17), který je posledním příkladem zapadajícím do systému, obsahuje dvě jmenné fráze, z nichž *klinická smrt* je označena jako termín a druhá, *svěcená voda*, je ve stejné složce s fokalizátorem *ani* (Graf 13).

29 SYNv8; *Deníky Bohemia*, 15. 9. 2004.
30 SYN2015 (Křen et al., 2015); Hill, T. (2004). *Kryptograf.* Překlad: Demlová, M. Praha: Egmont.
31 SYN2015; Petersová, E. (2004). *Neobyčejný benediktin*. Překlad: Pošustová, S. Praha: Mladá fronta.



GRAF 14: Složková struktura věty (18).

(17) *Je v **klinické smrti**, té už nepomůže ani **svěcená voda**.*[32]

Podobně jako u věty (10) s verbonominálním spojením *dát najevo* rozhodla gramatika i v tomto příkladu, že neodpovídá valenčnímu rámci. Zde ale nešlo o valenci víceslovného lexému, ale slovesa *pomoci*. Pád zájmena *té* byl totiž v předchozí analýze chybně určen jako genitiv.

Následují příklady, které odporují systému české syntaxe. V příkladu (18) je kromě víceslovné jednotky *mít (něco) za sebou*, která se nijak nevymyká z jazykového systému, použito spojení *neřád neřádská*.

(18) *Ten **má** určitě leccos **za sebou**, **neřád neřádská**.*[33]

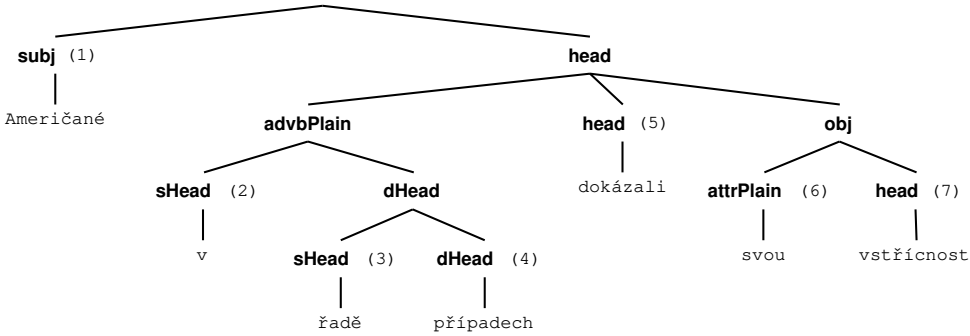
V případě věty (18) gramatika selhala a nevydala žádný výsledek (proto se také graf liší — nepochází z výstupu Tralu, ale z jeho vstupu). Důvodem je neshoda v rodu mezi slovem *neřád* a jeho adjektivním přívlastkem *neřádská*. Nicméně parser si s větou poradil, a dokonce je i ve frázi *neřád neřádská* správně určena hlava a přívlastek.

Ačkoli tato konstrukce odporuje české gramatice a je tedy nesystémová, je vžitá jako expresivní výraz, nebo spíše vzor, podle kterého se dají další výrazy tvořit (*kluk ušatá*, *chlap mizerná* atd.). Je tedy žádoucí v tomto případě uvolnit omezení daná gramatikou tak, aby věta nebyla označena za chybnou, ale u spojení *neřád neřádská* bylo vyznačeno, že není v souladu se systémem.

Poslední příklad ukazuje konstrukci, která se vymyká jazykovému systému a je i pociťována jako nesprávná, ale přesto se v textech vyskytuje. Můžeme pouze spekulovat, jak k této chybě dochází — zda po záměně slova *mnoha* za slovo *řadě* pisatel zapomene opravit tvar *případech* na správný tvar *případů*, nebo jestli má vliv předložka *v*, která vyžaduje lokálovou rekci, anebo jestli někteří mluvčí jazyka začínají pociťovat tvar *řadě* v této konstrukci spíše jako číslovku.

32 SYNv8; *Sport*, 15. 2. 2017.

33 SYNv8; Šukšin, V. M. (1987). *Červená kalina*. Překlad: Psůtková Z. Praha: Lidové nakladatelství.



GRAF 15: Složková struktura věty (19).

(19) *Američané **v řadě případech** dokázali svou vstřícnost.*[34]

Stejně jako v předchozím případě i u této věty gramatika nevydala žádný výsledek. Na rozdíl od předchozího případu ale nemá být gramatika opravena tak, aby nesystémovou konstrukci připustila, ale měla by na ni upozornit.

7. ZÁVĚRY

Implementovali jsme a otestovali lingvisticky motivovanou složkovou gramatiku založenou na pravidlech s cílem zkontrolovat a doplnit anotaci textů provedenou stochastickými nástroji. K tomu gramatika využívá valenční slovník a databázi víceslovných lemmat. Gramatika se interpretuje jako constraint solver, který považuje morfosyntakticky a strukturně anotované texty, pravidla gramatiky a lexikální hesla za množinu omezení, která si nemají odporovat. Výsledky budou využity pro vytvoření korpusu s ověřenou anotací. Součástí této anotace bude i informace, které konstrukce nebo tvary jsou v rozporu s obecnými gramatickými pravidly. Na jejím základě bude možné statisticky vyhodnotit všechny případy, kdy je úzus v rozporu s gramatikou, a to na nejrůznějších typech jevů.

Zde jsme se zaměřili na anotaci víceslovných lemmat v jedenácti větách, z nichž většina obsahuje alespoň dvě taková lemmata. Ukázali jsme, jak gramatika doplňuje údaje o víceslovných lemmatech do neterminálních uzlů a jak umožňuje kontrolu analýzy provedené stochastickými nástroji. I když se složková syntaktická struktura často liší od struktury, kterou bychom u víceslovných lexémů očekávali, ukázali jsme přednosti dvojího pohledu. Lexikální pohled je vyjádřen informacemi v uzlech stromu a syntaktický jeho strukturou.

I když jsme ukázali řadu typů víceslovných lexémů, některé typy (např. rčení) jsme nechali stranou kvůli zaměření na interakci lexémů se syntaktickým kontextem a z prostorových důvodů. Jde zejména o příklady syntaktické, morfologické

34 SYNv8; *Mladá fronta DNES*, 8. 7. 1999.

a lexikální idiomatičnosti, které se týkají např. neobvyklých valenčních vlastností (*Když Pán Bůh dopustí, i motyka spustí. Lehce nabyl, lehce pozbyl.*), elidovaného slovesa (*My o vlku, vlk za dveřmi. Někdo holky, jinej vdolky. Svůj k svému. Národ sobě.*) nebo jinak neužívaných slov v neobvyklé syntaktické struktuře (*láry fáry, třesky plesky, tintili vantili*).

Další vývoj gramatiky se zaměří dvěma směry: na vyšší pokrytí a přesnost kontrolovaných jevů a struktur a také na automatickou diagnostiku případů, kdy anotace nesplňuje předpoklady gramatiky.

SEZNAM LITERATURY

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, 2nd edition (s. 267–292). Boca Raton, FL: CRC Press.

Bejček, E., & Straňák, P. (2010). Annotation of multiword expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, *44*(1–2), 7–21.

Cvrček, V. (2013). *Kvantitativní analýza kontextu*. *Studie z korpusové lingvistiky*, svazek 18. Praha: Nakladatelství Lidové noviny.

Čermák, F., Hronek, J., Machač, J., Blatná, R., Churavý, M., Červená, V., Holub, J., Kopřivová, M., Kroupová, L., Mejstřík, V., Šára, M., & Trnková, A. (1983–2009). *Slovník české frazeologie a idiomatiky* (*SČFI*), svazek 1–4. Praha: Academia/Leda.

Dyvik, H., Losnegaard, G. S., & Rosén, V. (2019). Multiword expressions in an LFG grammar for Norwegian. In Y. Parmentier & J. Waszczuk (Eds.), *Representation and Parsing of Multiword Expressions: Current Trends* (s. 69–108). Berlin: Language Science Press.

Hnátková, M. (2002). Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*, *63*(2), 117–126.

Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., & Vondřička, P. (2017). Eye of a needle in a haystack. Multiword expressions in Czech: Typology and lexicon. In R. Mitkov (Ed.), *Computational and Corpus-Based Phraseology* (s. 160–175). Berlin: Springer.

Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., & Vondřička, P. (2018). Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. *Korpus — gramatika — axiologie*, *9*(17), 3–22.

Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., & Vondřička, P. (2019). Lexical database of multiword expressions in Czech. In V. P. Zakharov (Ed.), *Trudy meždunarodnoj konferencii „Korpusnaja lingvistika — 2019“* (s. 9–16). St. Petersburg: Saint Petersburg University Press.

Jelínek, T. (2016). Combining dependency parsers using error rates. In P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, Speech and Dialogue — Proceedings of the 19th International Conference TSD 2016* (s. 82–92). Berlin: Springer.

Jelínek, T., Petkevič, V., Rosen, A., Skoumalová, H., Vítovec, P., & Znamenáček, J. (2014). A grammar-licensed treebank of Czech. In V. Henrich, E. Hinrichs, D. de Kok, P. Osenova & A. Przepiórkowski (Eds.), *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT13)* (s. 218–229). Tübingen.

Jelínek, T., Petkevič, V., Rosen, A., Skoumalová, H., & Vítovec, P. (2015). Taking care of orphans: Ellipsis in dependency and constituency-based treebanks. In M. Dickinson, E. Hinrichs, A. Patejuk & A. Przepiórkowski (Eds.), *Proceedings of the Fourteenth International Workshop on Treebanks*

and Linguistic Theories (TLT14). (s. 119–133). Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Kettnerová, V., Lopatková, M., Bejček, E., & Barančíková, P. (2018). Enriching VALLEX with light verbs: From theory to data and back again. *The Prague Bulletin of Mathematical Linguistics*, 111, 21–56.

Kopřivová, M., & Hnátková, M. (2014). From dictionary to corpus. In V. Jesenšek & P. Grzybek (Eds.), *Phraseology in Dictionaries and Corpora, Proceedings of EUROPHRAS 2012*, Mednarodna knjižna zbirka ZORA 97 (s. 155–168). Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta, Univerza v Mariboru.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., & Zasina, A. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., & Zasina, A. (2019). *Korpus SYN*, verze 8 z 12. 12. 2019. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://www.korpus.cz.

Lopatková, M. (2015). *T-rovina PDT: Víceslovné výrazy (PDT 2.5, 3.0)*. Prezentace dostupná z https://ufal.mff.cuni.cz/~lopatkova/2015/docs/12-t-mwe.pdf.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., & Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Praha: Karolinum.

Markantonatou, S., Ramisch, C., Savary, A., & Vincze, V. (2018). *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*. Berlin: Language Science Press. Dostupné z https://langsci-press.org/catalog/book/204.

Parmentier, Y., & Waszczuk, J. (2019). *Representation and Parsing of Multiword Expressions: Current Trends*. Berlin: Language Science Press. Dostupné z https://langsci-press.org/catalog/book/202.

Petkevič, V., Rosen, A., Skoumalová, H., & Vítovec, P. (2015). Analytic morphology — merging the paradigmatic and syntagmatic perspective in a treebank. In J. Piskorski, L. Pivovarova, J. Šnajder, H. Tanev & R. Yangarber (Eds.), *The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)* (s. 9–16). Hissar, Bulgaria.

Pollard, C. J., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Przepiórkowski, A., Hajič, J., Hajnicz, E., & Urešová, Z. (2017). Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, *30*(1), 1–38.

Richter, F., & Sailer, M. (2014). Idiome mit phraseologisierten Teilsätzen: Eine Fallstudie zur Formalisierung von Konstruktionen im Rahmen der HPSG. In A. Lasch & A. Ziem (Eds.), *Grammatik als Netzwerk von Konstruktionen* (s. 291–312). Berlin: de Gruyter.

Rosen, A. (2018). Coping with unruly language: Non-standard usage in a corpus. In E. Fuß, M. Konopka B. Trawinski, & U. H. Waßner (Eds.), *Grammar and Corpora 2016* (s. 271–287). Heidelberg: Heidelberg University Publishing.

Rosen, A., & Skoumalová, H. (2018). No way to have your say out of the frame: Specifying valency of multi-word expressions. *Prace Filologiczne*, LXXII, 301–320.

Sailer, M., & Markantonatou, S. (2018). *Multiword expressions: Insights from a multilingual perspective*. Berlin: Language Science Press. Dostupné z http://langsci-press.org/catalog/book/184.

Sailer, M., & Richter, F. (2002). Not for love or money: Collocations! In G. Jäger, P. Monachesi, G. Penn & S. Wintner (Eds.), *Proceedings of Formal Grammar 2002* (s. 149–160). Trento, Italy.

Skoumalová, H. (2016). Slovníková pravidla pro využití slovníku VALLEX při tvorbě syntakticky anotovaného korpusu. In K. Skwarska & E. Kaczmarska (Eds.),

Výzkum slovesné valence ve slovanských zemích (s. 131–147). Praha: Slovanský ústav AV ČR.

Skoumalová, H., Hnátková, M., & Petkevič, V. (2011). Linguistic Annotation of Corpora in the Czech National Corpus. In V. Zacharov (Ed.), *Trudy meždunarodnoj konferencii „Korpusnaja lingvistika — 2011"* (s. 15–20). Sankt-Petěrburg: St.-Petersburg State University, Institute of Linguistic Studies.

Woliński, M. (2019). *Automatyczna analiza składnikowa języka polskiego*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Alexandr Rosen | Ústav teoretické a komputační lingvistiky FF UK
<alexandr.rosen@ff.cuni.cz>

Hana Skoumalová | Ústav teoretické a komputační lingvistiky FF UK
<hana.skoumalova@ff.cuni.cz >

Jiří Znamenáček | Ústav informatiky a chemie, Vysoká škola chemicko-technologická v Praze
<Jiri.Znamenacek@vscht.cz>