Neprojektivity v Pražském závislostním korpusu (PDT)

Technická zpráva CKL/ÚFAL, 2004

Daniel Zeman

Cílem této technické zprávy je popsat neprojektivní konstrukce v Pražském závislostním korpusu (Prague Dependency Treebank, PDT; Hajič et al., 2001), roztřídit je podle kategorií, ke každé kategorii uvést počty výskytů neprojektivit, které do ní patří, a příklady neprojektivit z PDT. Cílem zprávy není neprojektivity nijak vysvětlovat, srovnávat s hypotézami v různých teoriích apod. (k tomu viz např. Hajičová et al., 2004).

Výzkum, jehož výsledkem je tato zpráva, byl podporován projektem MŠMT č. LN00A063 (Centrum komputační lingvistiky).

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

Definice pojmů

Definice 1. *Věta* je posloupnost slov, přičemž za *slovo* se považuje i číslo nebo interpunkční znaménko (anglicky též *token*). *Index slova* je přirozené číslo, udávající pozici slova ve větě; první slovo má index 1, poslední slovo (obvykle koncová interpunkce) má index *N*, kde *N* je počet slov ve větě. *Slovosled* je úplné uspořádání slov ve větě, odpovídající relaci < nad množinou indexů slov v této větě.

Definice 2. *Závislostní struktura věty* neboli *strom* je dvojice (V, H) taková, že V je množina *uzlů (vrcholů)*, H je množina *závislostí (hran)* a dále platí:

1. Uzlů je o 1 více než slov v podkladové větě. Přidaný uzel má index 0, nazývá se *kořen* a neodpovídá žádnému slovu podkladové věty. Ostatní uzly jsou ohodnoceny slovy podkladové věty a mají stejné indexy jako slova.
2. Závislost je uspořádaná dvojice uzlů (*ř*, *z*), přičemž uzel (s indexem) *ř* nazýváme *řídící uzel* nebo *rodič*, uzel (s indexem) *z* nazýváme *závislý uzel* nebo *dítě*. Uzly mající společného rodiče nazýváme *sourozenci*. Říkáme též, že uzel *z* závisí na uzlu *ř*; tuto relaci značíme *ř*↓*z*, popř. ekvivalentně *z*↑*ř*.
3. Závislostí je právě tolik, kolik je slov v podkladové větě (*N*). Pro každé slovo *s* podkladové věty existuje právě jeden uzel $x \in \langle 0; N \rangle$ takový, že *x*↓*s*. Kořen je jediný uzel, který nezávisí na žádném jiném uzlu.
4. Relace ⇊ je tranzitivním uzávěrem relace ↓. Platí, že *x*⇊*y* právě tehdy, když buď *x*↓*y*, nebo existuje uzel *z* takový, že *x*↓*z* a *z*⇊*y*. Potom říkáme, že v dané závislostní struktuře je uzel *x* *předkem* uzlu *y* a uzel *y* je *potomkem* uzlu *x*.
5. Závislostní struktura neobsahuje *cykly*. To znamená, že pro žádný její uzel *x* neplatí *x*⇊*x*.
6. Uzly, které v dané závislostní struktuře nemají děti, se nazývají *listy*.

Definice 3. *Podstrom T_x uzlu x ve stromu T=(V, H)* je dvojice (V_x, H_x) taková, že platí:

1. $V_x \subseteq V$ a všechny uzly $y \in V$, pro které platí $x=y$ nebo $x \Downarrow y$, jsou prvky V_x.
2. $H_x \subseteq H$. H_x obsahuje právě takové závislosti $d=(ř, z)$, kde *ř* i *z* jsou prvky V_x.

Uzel *x* nazýváme *kořenem podstromu*.

Poznámka: Podstrom v našem smyslu tedy není libovolná podmnožina vyříznutá kdekoli ve stromu, ale vždy výhradně od určitého uzlu dolů až ke všem listům.

Definice 4. Závislost (*ř*, *z*) je ***neprojektivní***, jestliže existuje uzel *x*, který leží ve větě mezi *ř* a *z*, ale neleží v podstromu uzlu *ř*. Formálněji, $(ř < x < z) \wedge \neg(ř \downarrow\downarrow x)$ nebo $(ř > x > z) \wedge \neg(ř \downarrow\downarrow x)$. Strom je neprojektivní, jestliže obsahuje alespoň jednu neprojektivní závislost. Nebude-li v konkrétním případě řečeno jinak, *neprojektivitou* budeme mít na mysli neprojektivní závislost.

Definice 5. ***Díra*** (anglicky *gap*) je nepřerušená řada slov (uzlů) mezi řídícím a závislým uzlem neprojektivní závislosti (*ř*, *z*), takových, že pro žádný uzel *x* v díře neplatí *ř*⇊*x*.

Pozorování 1. Mějme závislostní strukturu zakreslenou jako diagram, kde uzly odpovídají bodům, závislosti spojnicím mezi body, vodorovné souřadnice uzlů vyjadřují jejich úplné uspořádání podle relace < a svislé souřadnice vyjadřují jejich částečné uspořádání podle relace ↓. (Viz též kteroukoli ilustraci u příkladů uvedených dále v této zprávě.) Proveďme *projekci* stromu na základnu tak, že z každého uzlu spustíme kolmici k základně (vodorovné čáře ležící níže než nejnižší list). Potom každá neprojektivní závislost bude protnuta nejméně jednou takovou kolmicí, zatímco projektivním závislostem se nic podobného nestane. Odtud pochází pojem *(ne)projektivita*.

Poznámka: Může se dokonce stát, že neprojektivní závislost protne jinou závislost, ale tato podmínka není nutná. V angličtině se někdy pro neprojektivitu vágně používá pojem *crossing dependency*, ale ten je vzhledem k právě uvedenému zavádějící.

Pozorování 2. Počet děr nemá přímou souvislost s počtem neprojektivit. Prostor mezi řídícím a závislým uzlem neprojektivní hrany může být rozdělen na několik děr. Naopak jednu díru může překlenovat několik neprojektivit (které mohou, ale nemusí sdílet stejný řídící uzel).

Celkové počty neprojektivit v PDT

V této zprávě se zabýváme pouze neprojektivitami na analytické rovině v PDT 1.0. K jejich identifikaci ovšem využíváme atributy uzlů, které pocházejí z různých rovin (morfologické, analytické a ojediněle i tektogramatické).

Veškeré statistiky pocházejí z té části dat, která je v PDT 1.0 označena jako trénovací data pro analytické parsery. Tato část obsahuje 1583 souborů (c*, l101–lt52, m*, v*), 73 088 neprázdných vět (z 81 614 vět celkem) a 1 255 590 slov (tokenů).

Neprojektivita je poměrně řídký jev vzhledem k počtu závislostí, ale relativně běžný jev vzhledem k počtu vět.

Celkem **23 691 závislostí (1,9 %)** bylo neprojektivních.

Celkem **16 920 vět (23,2 %)** obsahovalo alespoň jednu neprojektivitu.

Klasifikace neprojektivit

Vycházíme z třídění uvedeného v (Hajičová et al., 2004), které dále zjemňujeme a rozšiřujeme. V první instanci dělíme neprojektivity do tří tříd A, B, C.

A: Technické neprojektivity. Samy zmizí při přechodu na tektogramatickou rovinu. To se může stát v důsledku jednoho z následujících jevů:

1. Přestane existovat neprojektivní hrana, protože buď
    a. závislý uzel na TR zmizí a stane se pouhou vlastností řídícího, nebo
    b. řídící uzel zmizí a stane se vlastností závislého.
2. Zmizí díra, protože
    a. kořeny podstromů v díře zmizí a stanou se pouhou vlastností svých rodičů (příklad: *půjdeme-li vpravo*; v díře je pomlčka, ta se slepí s *li*, a dále *li* samo, to se přilepí k *půjdeme*), nebo
    b. řídící uzel neprojektivity zmizí a stane se vlastností kořene díry (pokud má díra více než jeden kořen, pak k tomu asi nedojde).

B: Analytické neprojektivity. Při přechodu na tektogramatickou rovinu nezmizí samovolně, ale budou potlačeny násilnou projektivizací — změnou slovosledu[1]. Ve všech případech by mělo jít o neprojektivity způsobené změnami slovosledu kvůli výpovědní dynamičnosti.

C: Některé typy frazémů. Výpovědní dynamičnost (konkrétně kontrast) zde není v seznamu viníků na prvním místě. Do C nespadají pouze neprojektivity, u kterých ke kontrastu nikdy nedochází, nýbrž takové typy, u kterých k němu *nemusí docházet ve všech případech*. Problémem je zejména to, že frazém reprezentuje sémanticky jeden objekt, ale rozpadl se do více uzlů. Tyto neprojektivity zůstanou i na tektogramatické rovině, pokud ovšem jednou nebudou všechny frazémy staženy do jediného uzlu.

[1] Jinou možností projektivizace, používanou např. u některých parserů, je posunutí závislosti směrem vzhůru, tj. řídícím uzlem se stane předek dosavadního řídícího uzlu.

Následuje přehled jednotlivých typů neprojektivit podle právě zavedených tříd. V některých případech bylo vymezení typů přizpůsobeno požadavku, aby bylo možné daný typ automaticky rozpoznat podle anotací, které jsou v PDT k dispozici. U každého typu podrobně popisujeme, podle čeho byl typ v PDT identifikován a čemu tedy odpovídají uvedené počty výskytů.

Využívají se hodnoty následujících atributů (u řídícího uzlu neprojektivity, u závislého uzlu nebo u uzlů v díře):

- lemma (heslo, <l> — např. u seznamu modálních sloves)
- morfologická značka (m-značka, tag, <t>)
- analytická funkce (s-značka, afun, <A>)
- tektogramatický funktor (kvůli hodnotám CPHR a DPHR; resp. technicky se to dělá tak, že se z dat na tektogramatické rovině vytáhne seznam všech slov, která mohou mít dotyčný funktor, a na analytické rovině už se pak pouze testuje, zda dotyčné slovo najdeme v seznamu pro příslušný funktor)

U některých příkladů zobrazujeme pro ilustraci jak jejich analytický strom, tak i odpovídající tektogramatický (analytický vždy vlevo popř. nahoře, tektogramatický vpravo popř. dole).

Tam, kde u příkladů odkazujeme na jejich výskyt v PDT, používáme místo identifikátoru věty dle ČNK (atribut id CSTS prvku <s>) dvojici jméno souboru : číslo věty v něm. Na rozdíl od identifikátoru ČNK existuje riziko, že v budoucích verzích PDT daný odkaz povede na jinou větu; na druhou stranu je ovšem mnohem snadnější vyhledat soubor, ve kterém se věta nachází, otevřít ho v Tredu a zobrazit větu s příslušným číslem. Příklad: „c101:4".

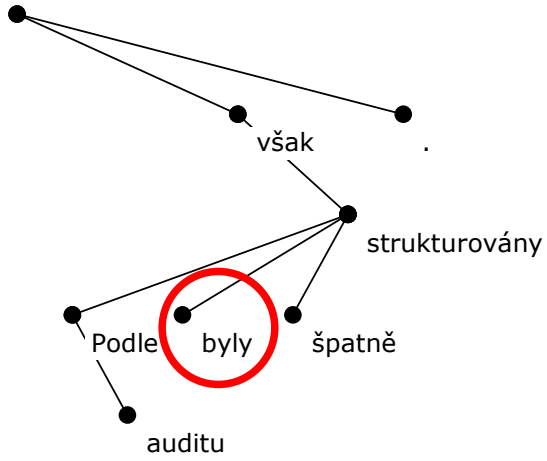
Uzly v neprojektivní hraně (případně další slova z jejich podstromu) jsou v příkladech zvýrazněny tučně. Tam, kde jsme to považovali za vhodné, je také podtrženo slovo v díře.

A1

Sem patří všechny neprojektivity, které zmizí, protože zmizí funkční slovo, které se na nich podílí (ať už jako vrchol neprojektivní hrany, nebo jako obsah díry).

A1-AuxV

Pomocná slovesa, zejména tvary slovesa *být*. Závislý uzel neprojektivit A1-AuxV má analytickou funkci AuxV. Jestliže neprojektivita patří do této kategorie, už nás nezajímá, zda také patří do jiných níže uvedených kategorií (a nebude zahrnuta do jejich statistik). Neprojektivit typu A1-AuxV bylo v datech nalezeno 402.



*Podle auditu **byly** však špatně **strukturovány**.* (c104:33)
Trpný rod. Slovo *byly* (*jsou*, *budou*) visí na příčestí trpném.

*Já **jsem** proto **předpokládal**, že…* (c228:16)
Minulý čas. Slovo *jsem* visí na příčestí minulém.

*S tím **by** však firma **neobstála**.* (ca16:36)
Podmiňovací způsob. Slovo *by* visí na příčestí.

*…, **chybět** jim však **budou** Kraut a Basta.* (la24:12)
Budoucí čas. Slovo *budou* visí na infinitivu. Jeho neprojektivita může být doprovázena neprojektivitou dalšího sousedního členu, ale také nemusí.

*…, což **by** mělo **být** učiněno.* (lb34:19)
Částice *by* visí na *být* místo na *mělo*. Já bych to tak nezavěsil, ale zdá se, že spíše než úlet anotátora je to anotační pravidlo PDT. Je to jediná neprojektivita, protože *což* je podmět a visí tedy na *mělo*.

*Hlediskem **by** mělo **být** zaměření kanceláře.* (cc13:30)
V řadě případů je AuxV doprovázena ještě další neprojektivitou, která s ní sdílí díru. Celá konfigurace je způsobena kontrastivním vytažením této další neprojektivity; AuxV ji musí doprovázet proto, že je to příklonka. Jak je doloženo výše, i AuxV může mít v díře některé z notorických děrovadel jako *však*, *sice*, *proto*.

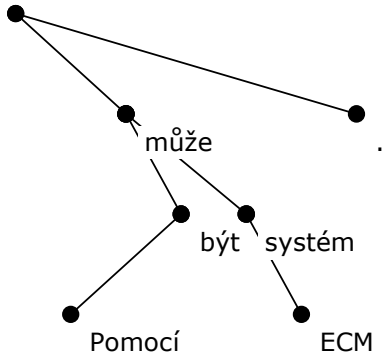
A1-modal

Viz též B7. Na rozdíl od ostatních konstrukcí s infinitivem, určitá tradičně vymezená skupina sloves přestane tvořit díru automaticky, protože zmizí a stane se pouhou vlastností infinitivu, který pod ní na AR visel.

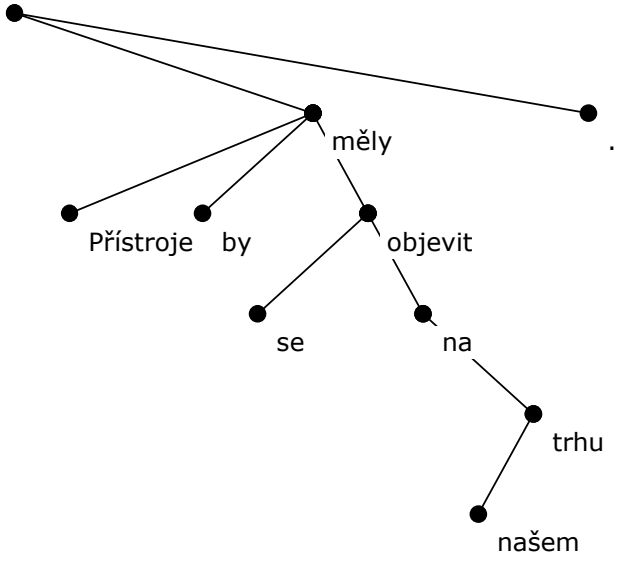
Zmíněné skupině se říká modální slovesa a dle (Hajičová et al., 2000) je vymezena následujícím seznamem:

muset, *mít*, *chtít*, *moci*, *dát se*, *smět*, *umět*, *dokázat*, *dovést*

Pokud je tedy řídícím uzlem neprojektivity infinitiv, ten závisí na slovesu z výše uvedeného seznamu a dotyčné modální sloveso je jediným prvkem díry, patří neprojektivita do kategorie A1-modal. Neprojektivit typu A1-modal se v datech našlo 3233.

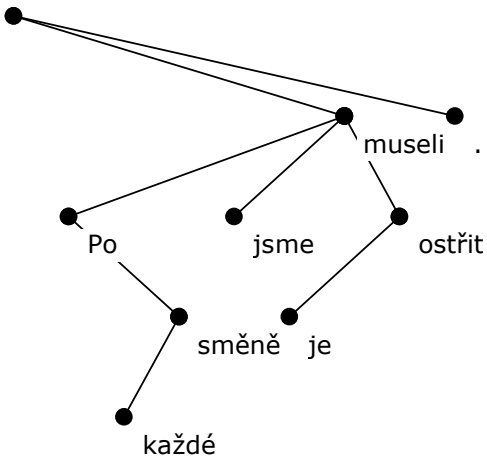


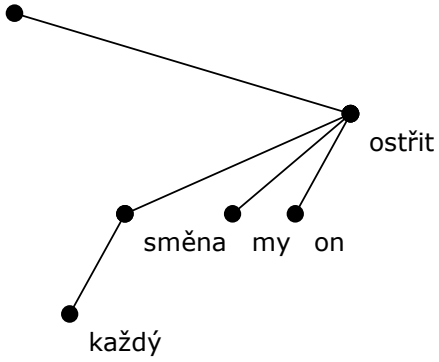
Pomocí *může* ***být*** *systém ECM.* (c101:4)



Přístroje by ***se*** *měly* ***objevit*** *na našem trhu.* (c101:45)

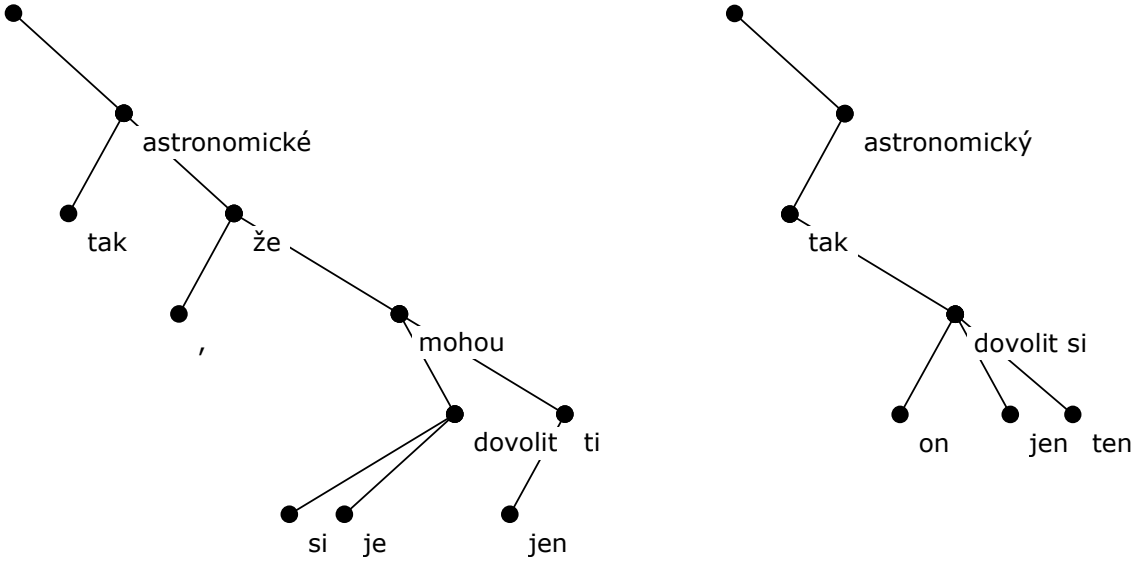
Po každé směně jsme ***je*** *museli* ***ostřit****.* (ca03:53)





*zboží, **kterým** může **nasytit** trh* (ca05:3)

> Mohlo by patřit do kategorie B6 (vztažná slova), ale nedojde k tomu, protože modální sloveso a s ním i neprojektivita přestane existovat.



*V rozhlase a televizi jsou ceny tak astronomické, že **si je** mohou **dovolit** jen ti, u nichž je kurz jejich měny vůči naší koruně výhodný.* (ca05:21)

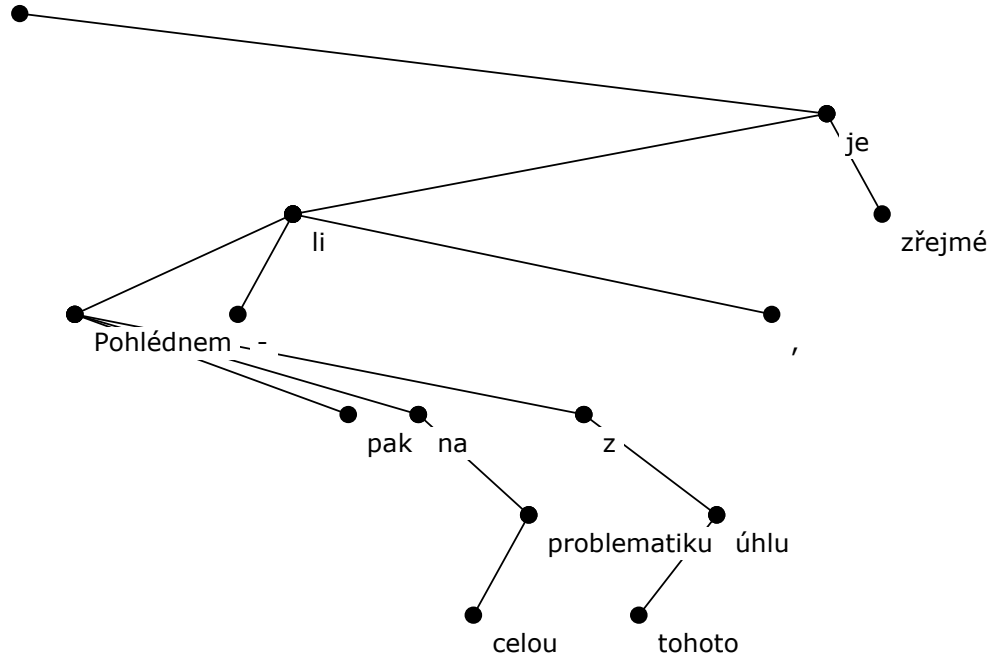
> Neprojektivita *dovolit-je* není zahrnuta do statistiky A1-modal, protože ji ještě předtím odchytila podskupina A1-AuxV. To musí být anotační chyba, funkce AuxV tam nemá co dělat, někdo asi omylem považoval zájmeno *je* za tvar slovesa *být* (přestože morfologická značka je přiřazena dobře, PPXP4—3). Ať tak či onak, obě neprojektivity samovolně zmizí spolu s modálním slovesem.
>
> Zajímavé pozorování na tektogramatické rovině: rematizátor *jen* se přesunul od zájmena *ti* pod sloveso *dovolit si*, ale hlavně o patro výš vznikla nová neprojektivita tím, jak se vedlejší věta přesunula pod *tak*!

A1-li

V díře je morfém *–li* s funkcí spojky podřadicí. Aby neprojektivita skutečně zmizela při přechodu na tektogramatickou rovinu, nesmí tam být už nic jiného (tedy kromě pomlčky, která visí na *li* a zmizí také).

Statistika nezahrnuje neprojektivity, které současně patří do kategorie A1-AuxV (např. *Budeme-li prodávat méně, bude nutné…*). Neprojektivit typu A1-li bylo v datech nalezeno 1089.

Poznámka: dalších 29 neprojektivit sice mělo v díře *li*, ale ještě něco k tomu, takže nezmizí samy od sebe. Někdy je tam navíc jen slovo *však*, *sice*, *ale*, *ovšem* apod. a *li* mimochodem není jen dírou, ale současně i uzlem zavěšeným kvůli dotyčnému slovu neprojektivně. Jindy pak jde o zcela jinou neprojektivitu, v níž se konstrukce *chcete-li* objevila náhodou (třeba vsuvka) a sloveso visící pod *li* není řídícím uzlem neprojektivity.

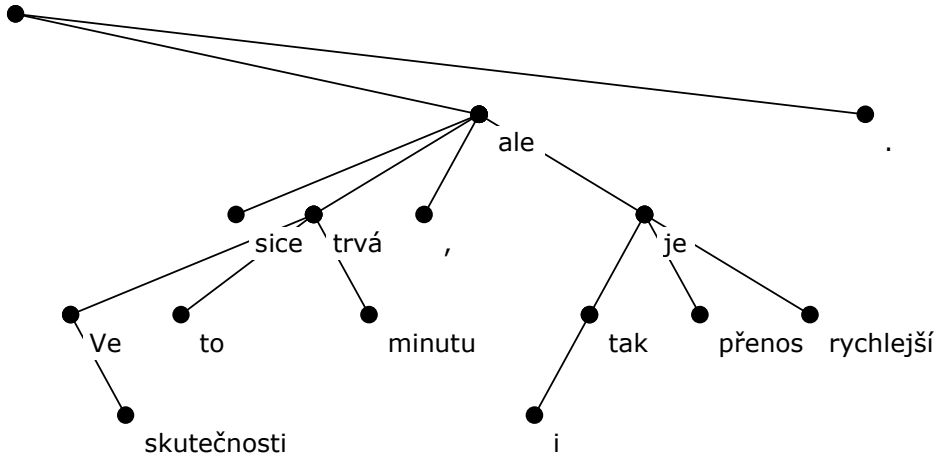


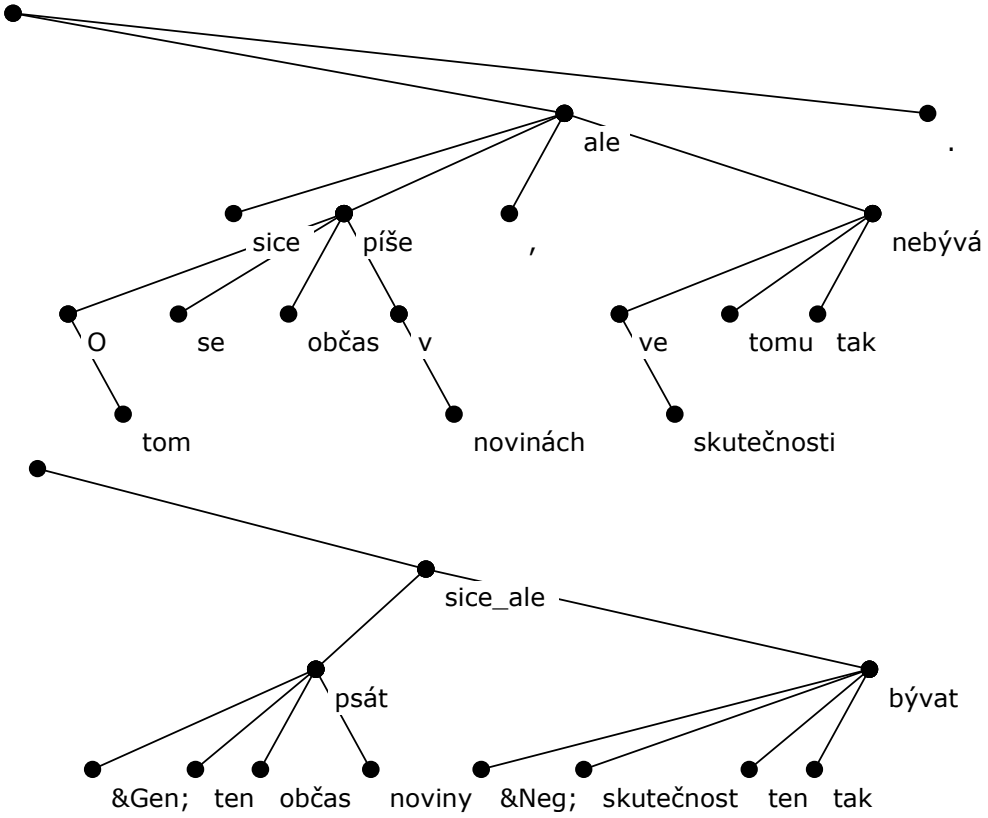
***Pohlédnem**-li **pak na celou problematiku z tohoto úhlu**, je zřejmé…* (ca05:21)

A1-sice-ale

Neprojektivita způsobená složeným spojkovým výrazem *sice – ale* (popř. *sice – avšak* atd.) V díře se nachází slovo *sice*, které visí na kořeni koordinace (afun Coord). Tento typ připomíná typ B1. Na rozdíl od něj však na tektogramatické rovině samovolně zmizí, protože uzel *sice* ztratí samostatnost. V PDT se celkem vyskytlo 407 neprojektivit tohoto typu.

Ve skutečnosti to sice trvá minutu, ale i tak je přenos rychlejší. (c101:9) Není anotován na tektogramatické rovině, nelze zkontrolovat.





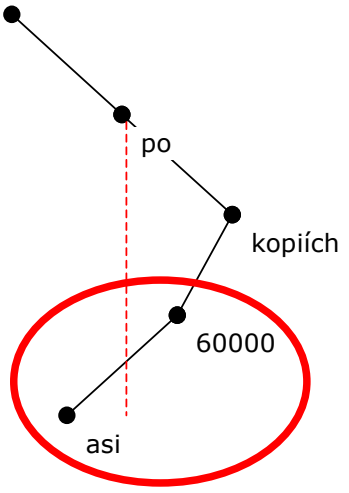
O tom se sice občas píše v novinách, ale ve skutečnosti tomu tak nebývá. (ca10:20)

A2: předložkové fráze (nejen) s rematizátorem

V díře je předložka (tedy uzel s morfologickou značkou začínající na R nebo jeden či více uzlů s analytickou funkcí začínající na AuxP), ale nic jiného. Kdyby tam bylo něco jiného, neprojektivita by nemusela samovolně zmizet při přechodu na tektogramatickou rovinu.

Neprojektivit typu A2-předložka bylo v datech nalezeno 6124, z toho 5850krát visel řídící uzel neprojektivní hrany přímo na předložce a 274krát jí byl podřízen nepřímo.

Příklad: *až k nečitelnosti*. Rematizátor *až* visí na podstatném jménu *nečitelnosti*, to celé visí na předložce *k*.



Příklady s nepřímou závislostí na předložce:

***asi** po **60000** kopiích* (c101:43)

***i** na **našem** trhu* (c101:45)

*čím větší obtíže budou **mít** naši klienti, s **tím** vyššími náklady bychom se museli potýkat* (c105:3)

> Vedlejší věta se dá asi stěží nazvat rematizátorem ("focus-sensitive particle"), ale to nic nemění na faktu, že při přechodu na tektogramatickou rovinu zmizí předložka a s ní i neprojektivita.

Do kategorie A2 patří i předložkové skupiny, jejichž řídící uzel není morfologická předložka (morfologická značka nezačíná na R), ale libovolný uzel s analytickou funkcí AuxP (syntaktická předložka). Syntaktická předložka se dokonce může skládat z několika slov.

*a **to** bez ohledu na **to**, že…*

teprve počátkem srpna (c139:47)

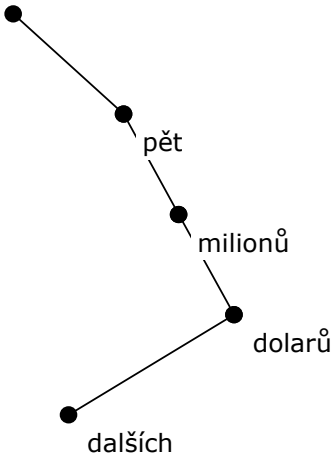
, a to včetně motoru (c212:44)

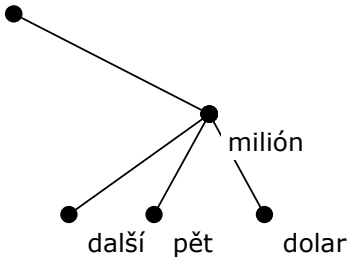
Následuje přehled nejčastějších slov, která se v neprojektivitách typu A2 vyskytla jako závislý uzel, tedy obvykle v pozici rematizátoru: *i* (1373), *až* (719), *jen* (441), *pouze* (278), *již* (252), *už* (227), *především* (222), *ještě* (209), *zejména* (200), *ani* (166), *právě* (158), *také* (155), *například* (148), *nejen* (116), *asi* (106), *třeba* (95), *přímo* (83), *to* (80), *hned* (66), *hlavně* (62), *teprve* (55), *např* (52), *alespoň* (51)…

A3: podstatné jméno závisí na číslovce

Neprojektivit typu A3-číslovka bylo v datech nalezeno 140.

Základní verze: Díra obsahuje jediný uzel, a to číslovku (morfologická značka začíná na C). Řídící uzel neprojektivity visí na této číslovce (jde obvykle o počítané podstatné jméno). Neprojektivita zmizí, protože na tektogramatické rovině bude číslovka považována za sémantické přídavné jméno a pověsí se naopak pod počítané podstatné jméno. (Takto se některé číslovky umísťují už na analytické rovině, ale jen pokud se





s počítaným podstatným jménem shodují v pádě.)

Příklady: *necelých dvacet haléřů*, *posledních deset let*, *příjemných 24 o* (stupňů), *Kč 125 tisíc*, *dalších pár lidí*, *z nichž pět tisíc*, *nás deset miliónů*.

Existují i složitější konstrukce, než výše uvedené, které sem patří.

dalších pět milionů dolarů (ca04:20)

V díře jsou tedy dvě slova a jen jedno z nich je označeno jako číslovka (*milión* je podstatné jméno). Objevily se pouze 3 výskyty takové konstrukce. Druhá:

z toho 200 miliónů korun připadne… (li43:56)

AR: připadne (200 (miliónů (korun (z (toho)))))

TR: Nelze zkontrolovat, protože soubor li43 není tektogramaticky anotován.

rekordních 2,3 milionu liber (lq34:10)

AR: 2,3 (milionu (liber (rekordních)))

TR: milión (rekordní, 2,3, libra)

Koordinace číslovek

*Zákaznice je pojištěna pro případ smrti na **Kč** 130 – 150 **tisíc** a na dožití na **Kč** 125 **tisíc**.* (cb19:10)

AR1: – (130, 150, tisíc (Kč))

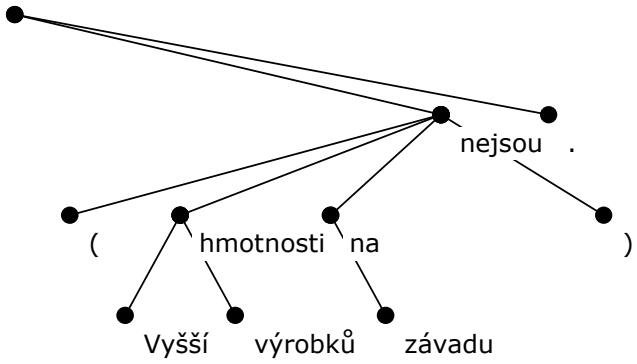
TR1: tisíc (Kč, &Hyphen; (130, 150))

AR2: 125 (tisíc (Kč))

TR2: tisíc (Kč, 125)

A4: AuxK – uzávorkované věty

Pokud je celá věta v závorkách nebo něčem podobném (v uvozovkách nebo třeba ve svislítkách), neprojektivitu způsobuje tečka na konci věty, která visí na kořeni, zatímco uzavírací závorka za ní visí na hlavním slovese a to je pod kořenem. Tato kuriozita není ani moc vzácná: našlo se jich 977!



*(Vyšší hmotnosti výrobků na závadu **nejsou.)*** (c111:33)

*| Kontakt**:** Ekonomická fakulta v Chebu, Hradební 22, 350 01 Cheb.* **|** (c120:47)

*… jde o to "vysoko si **vyhrnout** rukávy.***"*** (c210:50)

Na tomto příkladu je vidět, že „uzávorkovaná" nemusí být celá věta, stačí i poslední klauze.

A5: neprojektivně zavěšená interpunkce a výplňová slova

Sem řadím všechny neprojektivní závislosti, kde závislý uzel má analytickou funkci AuxG, AuxX, AuxY nebo AuxZ. Předpokládám, že tyto uzly na tektogramatické rovině nebudou a příslušná slova se nanejvýš stanou atributy jiných uzlů. V PDT se celkem vyskytlo 609 neprojektivit tohoto typu.

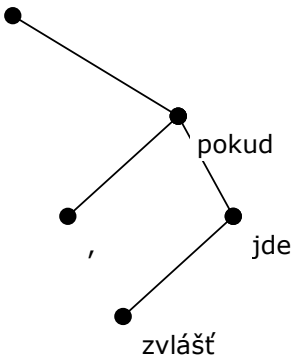
K neprojektivnímu zavěšení čárky (AuxX) dochází např. u spojení *stejně jako*:

*Kniha je určena pro zaměstnance poradenských firem**, stejně* ***jako*** *pro studenty ekonomických škol.* (c122:42)

AuxG: zatím jsem našel příklad, kde „věta" byla ve skutečnosti bodem seznamu, začínala na „b)", pak byla spojka *jestliže*, a pak teprve sloveso, na kterém viselo to „b)". ca22:33

AuxY: *a sice že prochází* (ca28:8).

Aux?: podobně jako AuxY u různých složených spojkových výrazů. *jen aby* (cd03:18).

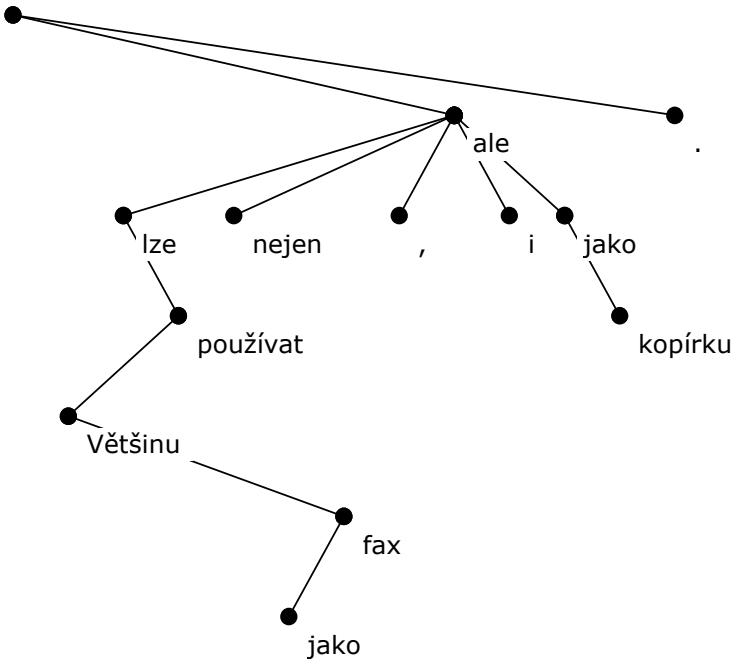


AuxZ: *, zvlášť pokud jde* (cc10:20). Ovšem v další větě (cc25:24) je více méně tentýž případ analyzován ne jako AuxZ, ale jako Adv!

B1: koordinace

Neprojektivit typu B1-koordinace bylo v datech nalezeno 835.

Uzel, který rozvíjí celou koordinaci, je zamotán v podstromu některého člena koordinace. Přesnější specifikace: V díře se nachází slovo, které visí na kořeni koordinace (afun Coord), ale samo není členem koordinace. Toto slovo nesmí být první díl složeného spojkového výrazu *sice – ale*, *nejen – ale*. Některé takové neprojektivity už sice zachytila kategorie A1, ale není to zaručeno. Pokud kromě prvního dílu spojky byla v díře ještě další slova, může jít o neprojektivitu, která při přechodu na tektogramatickou rovinu



nezmizí sama od sebe. Taková neprojektivita může patřit do jedné z dalších kategorií.

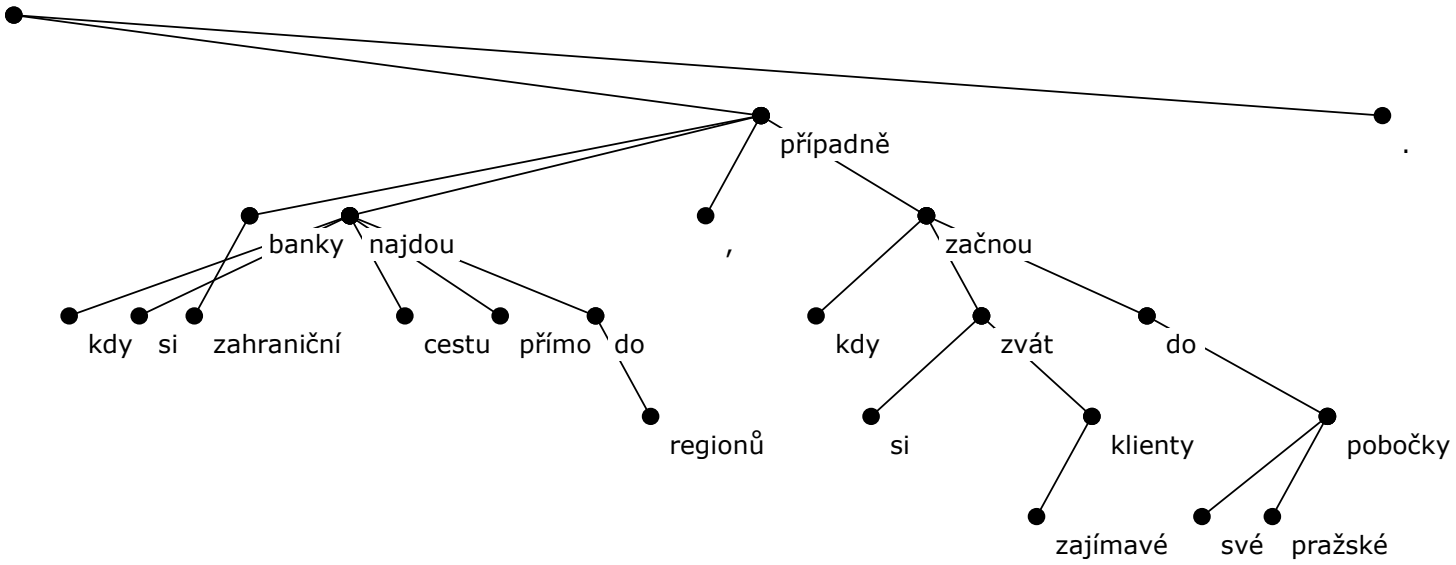
Příkladem neprojektivity, která z tohoto důvodu *nepatří* do B1, je následující věta:

Většinu *lze používat nejen jako* ***fax****, ale i jako kopírku.* (c101)

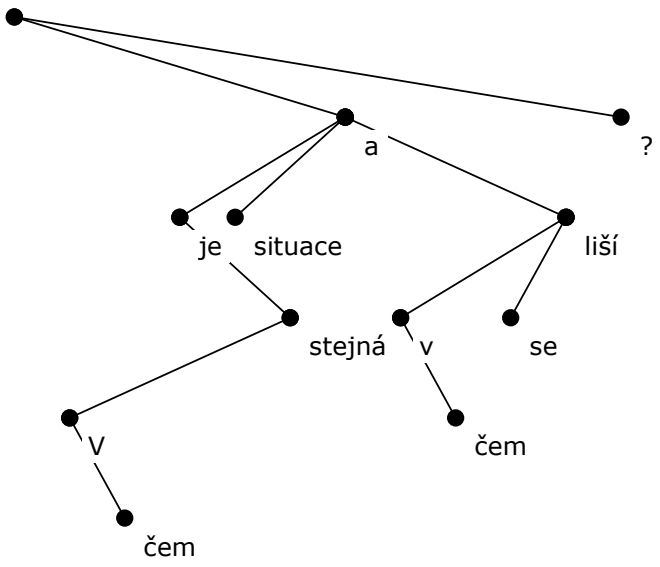
> Spojkový výraz *nejen – ale* se na tektogramatické rovině spojí, takže slovo *nejen* z díry zmizí, pořád v ní ale zůstává *lze používat*.

Z kategorie B1 vylučujeme také neprojektivity, které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

Skutečným příkladem B1 je následující věta:



…, ***kdy si*** *zahraniční banky* ***najdou*** *cestu přímo do regionů, případně kdy si začnou zvát*



zajímavé klienty do své pražské pobočky. (c105:8)
Klauze jsou koordinovány spojkou *případně*, *banky* jsou společným podmětem celé

koordinace, a visí tedy na *případně*, slova *kdy* a *si* jsou tím pádem oddělena od svého rodiče *najdou*.

V čem je *situace* ***stejná*** *a v čem se liší?* (c108:21)

Člověk ***si*** *již těžko* ***zvyká*** *a hledá nové známé.* (c114:16)
Slovo *těžko* visí na celé koordinaci, slova *si* a *již* pouze na slovese *zvyká* (*již* by možná mohlo také viset na celé koordinaci, ne tak *si* — i když se na první pohled zdá, že lze říct i *hledat si známé*, *si* u *zvykat* má jiný afun: AuxT).

B2: rozdělené frazémy typu DPHR

DPHR je funktor, kterým se dotyčné frazémy označují na tektogramatické rovině. Ještě tam existují frazémy typu CPHR, které odpovídají naší kategorii C1. U DPHR se počítá s tím, že v budoucnosti budou na tektogramatické rovině reprezentovány jediným uzlem. U CPHR je něco takového těžko představitelné, protože obsahují valenční podstatná jména (např. *zájem o něco*). Typický DPHR frazém obsahuje sloveso a 1 až 3 další slova (pokud je jich více, jde typicky o předložkovou vazbu). Sloveso má na tektogramatické rovině normální funktor (třeba PRED), slovo na něm závislé má funktor DPHR. Pokud frazém způsobuje neprojektivitu, pak je to tím, že se skládá z více uzlů. Jeho rozvití visí na závislém uzlu, ale je od něj odděleno dírou, ve které vězí hlava frazému – sloveso. Takové neprojektivity zmizí, až bude na tektogramatické rovině provedeno slití DPHR frazémů do jediného uzlu. Pak také budou neprojektivity typu B2 patřit spíše do třídy A. Momentálně ještě tektogramatická rovina není tak daleko, ale zmíněné neprojektivity se na ní opravují. Rozvití frazémů DPHR se převěšuje na hlavu frazému (sloveso), takže přestává být neprojektivní.

Máme k dispozici seznam DPHR frazémů, určený pro anotátory tektogramatické roviny. I když je tento seznam v principu otevřený, je to jediné vodítko, podle kterého můžeme rozpoznat frazémy strojově. Ukazuje se také, že je velmi těžké najít vůbec nějaké případy, kdy rozvití DPHR frazému bylo zavěšeno neprojektivně a kdy tato neprojektivita byla způsobena právě tím, že se frazém skládá z několika uzlů. Níže uvedený příklad, který byl v původní verzi našeho článku na Coling, totiž neobsahuje DPHR, ale CPHR frazém:

Se zuby jsem měl v minulosti jen problémy.

Frazém je *problémy se zuby*, *se zuby* visí na *problémy*.

Neprojektivit, které obsahovaly DPHR frazém a současně nesplňovaly definici neprojektivit B6 nebo B8, se našlo 20.

Musím pak ***na ně*** *dávat* ***pozor****.* (c137:4)

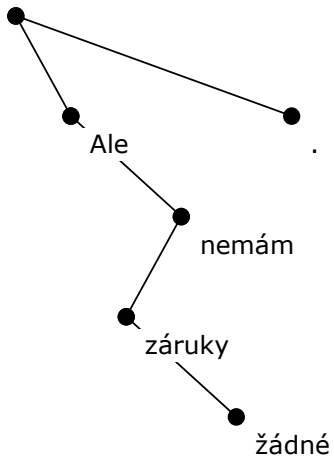
B3: rozdělené substantivní skupiny

Řídící uzel neprojektivity je podstatné jméno (jeho morfologická značka začíná na N), závislý uzel je jeho přívlastek. Co je v díře, nás tentokrát nezajímá.

Přívlastek poznáme podle toho, že analytická funkce závislého uzlu začíná na Atr. Zahrnujeme tím do B3 i neprojektivně zavěšené uzly s funkcí AtrAdv, což asi nevadí. Dále zahrneme i předložkové skupiny (neprojektivně zavěšený uzel má funkci AuxP, ale jeho nejlevější – a předpokládáme, že jediné – dítě má funkci začínající na Atr). Obdobně „průhledná" jako AuxP je pro nás i funkce AuxC, ale možná se to v tomto případě nijak neprojeví. A konečně zahrneme i koordinace a apozice přívlastků, resp. takové konstrukce, kde neprojektivně zavěšený uzel má funkci začínající na Coord nebo Apos, ale jeho nejlevějšímu dítěti, který má funkci končící na _Co, resp. _Ap, tato funkce začíná na Atr. U složených koordinací, resp. apozic, se sestup k dítěti rekurzivně opakuje.

Z kategorie B2/3 vylučujeme takové neprojektivity, které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

Neprojektivit typu B3, popř. B2 se v datech našlo 572. V tomto počtu nejsou zahrnuté doplňky, viz níže.



Výdaj to je dost velký.

*Ale **záruky** nemám **žádné**.* (c106:8)

Klasický příklad à la *Sportovec on je dobrý.*

*Podal jsem **žalobu** u soudu, **kterou jsem si sám napsal**.*

*Neboť jaký blázen by měl **zájem** kolem roku 1895 **padělat** holandského umělce, jehož plátna byla k dostání po dvou stech francích?*

Skupina *kolem roku 1895* je v díře, protože visí na *měl*, je však otázka, zda nemá spíše viset na *padělat*.

*Únorové komplikace kolem celní unie jsou **toho** v souvislosti se stavem slovenské platební bilance jen **dokladem**.*

***Průpravu** jsem měl **všeho druhu**.*

*mnoho **vozů**, možná většina, které byly **prodány***

Příčinou neprojektivity je vsuvka, což se stává často a myslím i u jiných kategorií. Vsuvky by si zasloužily samostatné zkoumání.

***S ní** neměli dosud **zkušenosti**.* (c112:46)

Závislý uzel (*s*) má funkci AuxP, ale jeho dítě (*ní*) má funkci Atr.

***O nábytek, doplňky a další řemeslné výrobky ze dřeva** byl vždy **zájem**.* (c129:28)

Tady je předložková konstrukce ještě navíc komplikována koordinací.

B3 – číslovky a množství

Významnou podskupinou B3 jsou spojení, která silně připomínají neprojektivity s číslovkami (viz A3 a B4), ale nesplňují definice příslušných skupin. Jejich společným rysem je, že jméno v genitivu visí na podstatném jméně vyjadřujícím množství. Skutečná číslovka se často objeví alespoň v díře. Skupina podstatných jmen (popř. příslovcí) vyjadřujících množství možná není uzavřená. V datech byla nalezena následující slova:

a) téměř číslovky: *desítka*, *stovka*, *sto*, *tisíc*, *milión*, *miliarda*

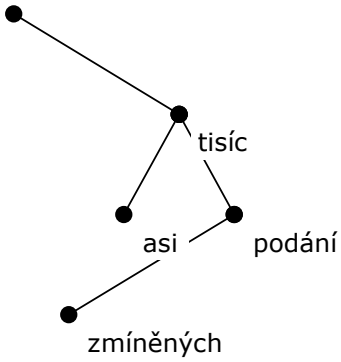
b) zlomky a násobky: *procento*, *polovina*, *dvojnásobek*, *párek*

c) neurčité „číslovky": *nedostatek*, *množství*, *nadbytek*, *menšina*, *dostatek*, *řada*, *většina*, *minimum*

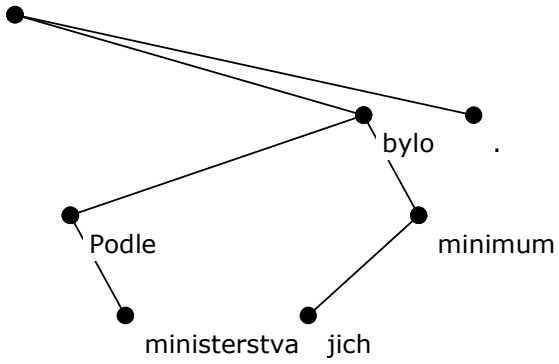
d) jednotky a jiné výrazy stojící mezi číslovkou a počítaným předmětem[2]: *tuna*, *hektolitr*, *druh*, *metr*

Pozor! Nemám jistotu, co se s těmito neprojektivitami stane na tektogramatické rovině. Zkoušel jsem se dívat do nepříliš staré verze tektogramatických dat, kterou jsem od Petra Pajase dostal někdy začátkem roku 2004, ale v ní jsou tyto konstrukce ponechány neprojektivní! Nebo se budou neprojektivity na tektogramatické rovině likvidovat až později? Nebo se to bude dělat jen ve vzorovém souboru?

Příklady:

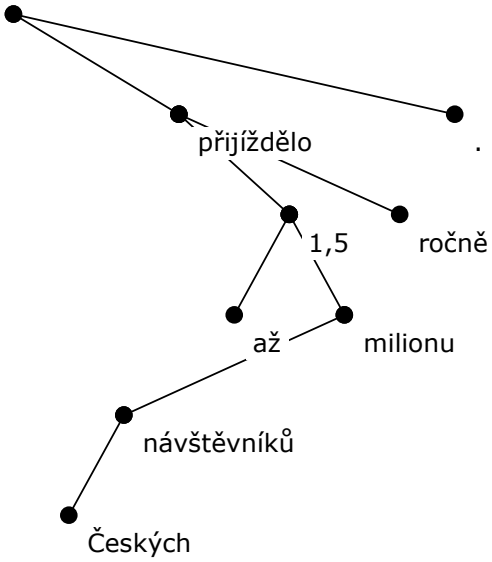


zmíněných asi tisíc podání (c103:47)

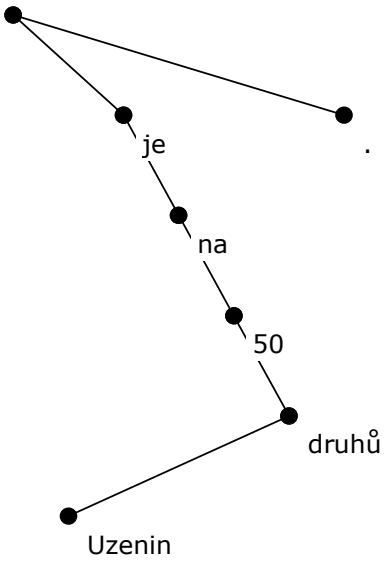


Podle ministerstva jich bylo minimum. (c126:25)

[2] Přísně lingvisticky vzato je ve výrazu *deset tun uhlí* počítaným předmětem samozřejmě *tuna* a ne *uhlí*, ale tady mi jde o samotný konec řetězce vztahů, o věc, jejíž množství určuje celá fráze.



Českých návštěvníků přijíždělo až 1,5 milionu ročně. (c205:17)



Uzenin je na 50 druhů. (c225:3)

celého půl milionu korun (ca08:38)

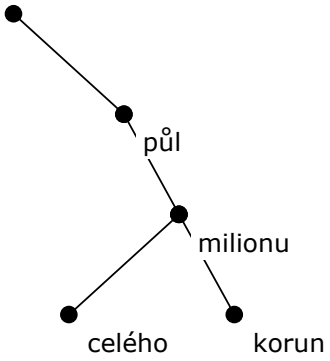
AR: půl (milionu (celého, korun))

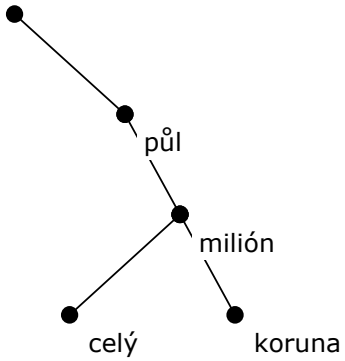
TR: půl (milión (celý, koruna))

Tedy žádná změna při přechodu z AR na TR! Já bych ovšem dal přednost stromu koruna (milion (půl (celý))).

Důvodů *je jistě celá* ***řada****.* (cb22:47)

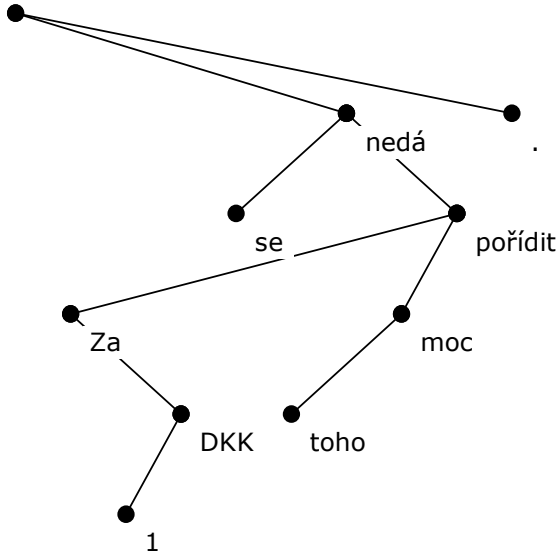
A tak dále. Mezi neprojektivitami typu B3 se celkem našlo 134 takových, kde závislý uzel neprojektivní hrany byl v genitivu, ne všechny však vyjadřují množství (viz třeba výše uvedený příklad *Únorové komplikace jsou toho jen dokladem.*) Ze statistiky lemmat řídících uzlů takových neprojektivit nicméně vysvítá, že podstatná část množství vyjadřuje. Nejčastějšími lemmaty bylo 7× *milión* a *tisíc* a 6× *procento* a *řada*.





Zvláštnosti:

*Za 1 DKK se **toho** nedá **moc** pořídit.* (cd09:41)



Dvě neprojektivity (pořídit-za a moc-toho) a nesouvislá díra (rozdělená slovem *toho*), jedna neprojektivita využívá obě části díry, jedna jen jednu část. Neprojektivita pořídit-za patří do kategorie B7 (infinitivy). Neprojektivita moc-toho do kategorie B7 nepatří, protože infinitiv není jejím řídícím uzlem. Ve skutečnosti ani nelze říct, že infinitiv je i její příčinou, protože totéž by se mohlo objevit i bez infinitivu: *Za 1 DKK se toho nepořídí moc.*

Jde tedy opět o neprojektivitu vyjadřující množství, v tomto případě by se však vůbec neměla objevit v B3, protože řídící uzel *moc* není podstatné jméno, ale příslovce nebo neurčitá číslovka. Chybou v morfologickém značkování však uzel dostal značku NNFS4 (ve významu *vláda*, *síla*).

B3 – doplněk

Velkou podskupinou či samostatnou skupinou jsou zřejmě **doplňky**. Od výše uvedené definice hlavní části skupiny B3 se liší tím, že analytická funkce neprojektivně zavěšeného uzlu nezačíná na Atr, ale na Atv (popř. AtvV). Kromě toho už ani nepožaduji, aby řídícím uzlem neprojektivity bylo podstatné jméno. Může to být tedy cokoliv, třeba vztažné zájmeno (*nezávislými podnikateli, kteří budou působit jako naši distributoři*) nebo, v případě AtvV, sloveso.

Do doplňků dále zahrneme i podstromy, kde je funkce Atv hlouběji skryta, protože jde o koordinaci doplňků (a kořen podstromu má tedy funkci Coord), popř. o apozici, předložkovou skupinu nebo skupinu řízenou podřadící spojkou.

V tektogramatických datech jsou doplňkové neprojektivity zlikvidovány zavěšením doplňku na sloveso.

Z kategorie B3 vylučujeme takové neprojektivity, které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

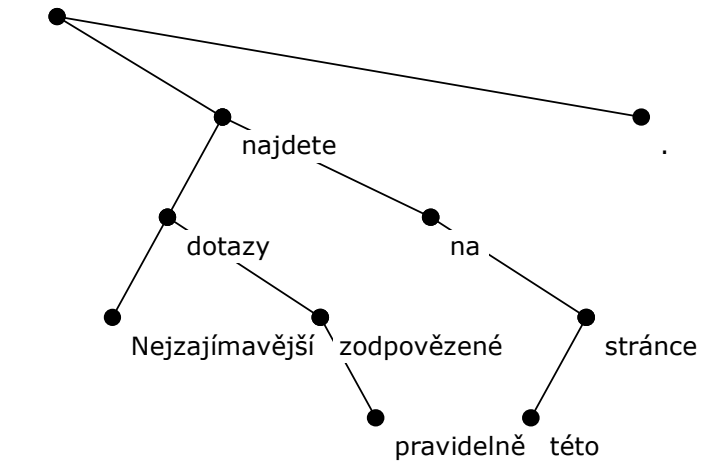
Neprojektivit typu B3-doplněk se v datech našlo 947 (toto číslo jsem nezahrnul do statistiky skupiny B3 uvedené výše, protože definice doplňkové podskupiny se dost liší).

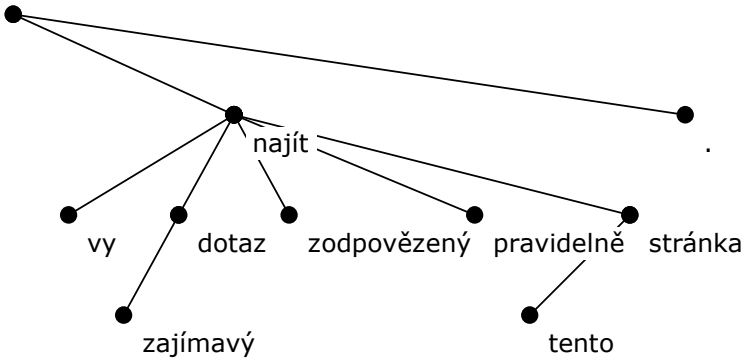
*… proč je nutné tento **parametr** brát **jen jako orientační údaj**.* (c101:12)

*firmy, **které** faxy **samy** dodávají* (c102:38)

*Jako další **úkol** jste si stanovili **revizi**.* (c105:17)

*nezávislými podnikateli, **kteří** budou působit **jako naši distributoři*** (c106:15)





*Nejzajímavější **dotazy** najdete **zodpovězené** pravidelně na této stránce.* (c110:46 a také ca02:38)

*Jeho **ražba** byla zahájená **jako odpověď** na záplavu českých grošů.* (c118:31)

***Obor** výpočetní techniky <u>je ke schvalování vyhlášen</u> **celý**.* (ca03:3)

AR: je (obor (techniky (výpočetní), celý), vyhlášen (ke (schvalování)))

TR: vyhlásit (obor (technika (výpočetní)), schvalování, celý)

*…, **které** jsou do prodeje uváděny **jako potraviny a nápoje*** (c131:46)

Příklad koordinace doplňků.

…, ***samotná*** *<u>však</u>* ***měla*** *pohledávky za 2,6 miliardy korun* (c104:32)

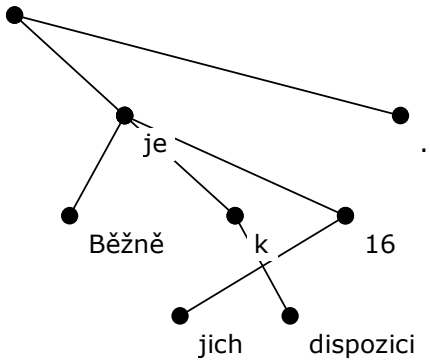
Zvážit, zda tento případ nevykázat do skupiny B8. Splňuje všechny její předpoklady, zatímco tady působí divně — doplněk totiž visí na slovese a jeho neprojektivita není způsobena nevyjasněností, zda má viset na podmětu nebo na přísudku (podmět ostatně není vyjádřen).

B4: číslovky oddělené od počítaných předmětů

Na tektogramatické rovině se sice obrátí směr závislosti, ale nic se nezmění na tom, že její koncové body jsou odděleny nesouvisející skupinou. Statistika zahrnuje neprojektivity, jejichž řídícím uzlem je na analytické rovině číslovka (morfologická značka začíná na C).

Z kategorie B4 vylučujeme takové neprojektivity, které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

Neprojektivit typu B4 se v datech našlo 265.



Běžně je ***jich*** *k dispozici* ***16****.* (c101:49)

Poutačů *je teď* ***tolik****, že část z nich ztrácí účinnost.*

Přenosová rychlost ***(A4/s)****<u>:</u>* ***15*** (c114:51)

Takovou větu jsme jistě nečekali. Vsuvka *(A4/s)* visí na *15*, v díře je dvojtečka – kořen apozice. Neprojektivity jsou hned 3, na *15* totiž visí samostatně *A* i obě závorky. Příklad není úplně ojedinělý, hned za ním se třeba našlo *Cena (Kč): 24900*.

Ze *stálého počtu potřebných třiceti pracovníků <u>se</u> jich asi* ***deset*** *vystřídalo.* (c129:34)

I tohle je příklad z jiného těsta, než jsme zamýšleli. Neprojektivní závislost na číslovce se netýká počítaného předmětu, ale omezující vlastnosti, navíc díru tvoří příklonka.

Z nich je jen jedna na kvalitu. (c224:16)

Z obrázků*, které visí na zdech,* ***jich*** *je v komisi asi* ***30****%.* (ca19:43)

Dvě neprojektivity, jednou omezující vlastnost, jednou počítaný předmět.

… před nímž kapituluje ***jedna*** *západní vláda* ***za druhou****.* (ca28:17)

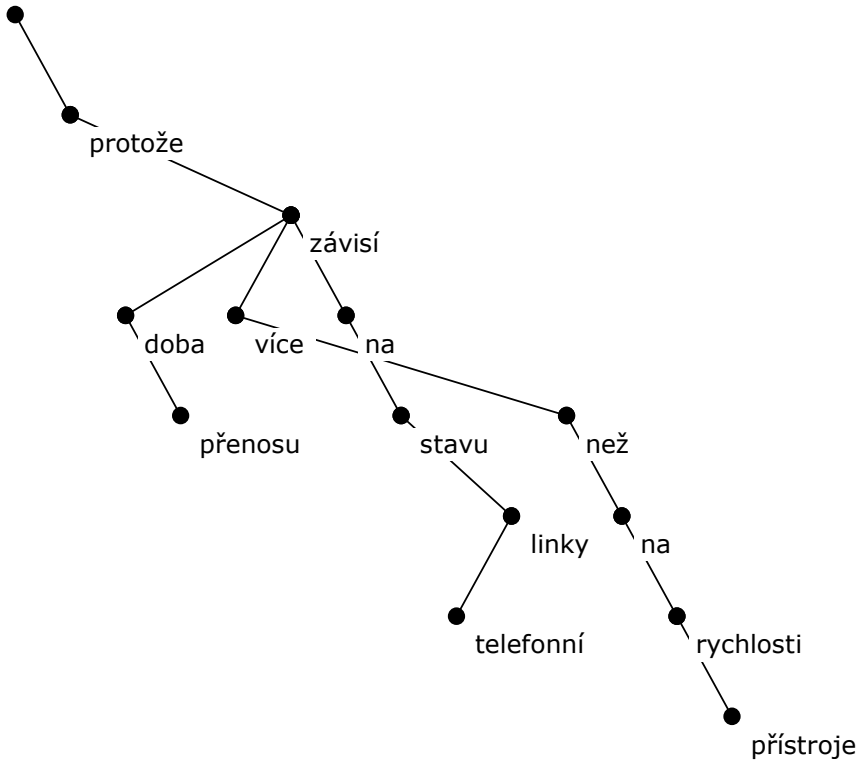
B5: rozdělené adjektivní či adverbiální skupiny srovnávací povahy

Typický příklad je dvojice *větší – než*, *více – než*, ale nemusí jít vždy o morfologický komparativ. Patří sem totiž i dvojice *stejný – jako*, *největší – na světě* apod. Protože tím

pádem není čeho se chytit, zkusím do B5 zahrnout všechny neprojektivity, jejichž řídící člen je přídavné jméno nebo příslovce, a pak uvidíme, co všechno do takové skupiny spadne.

Zpřesněné vymezení kategorie: Závislý uzel nesmí být vztažné slovo, protože v tom případě neprojektivita patří do kategorie B6 (*nejvyšší rychlostí, jaké je přístroj schopen*). Také nesmí jít o neprojektivní spojení člena apozice s kořenem apozice (samostatná nově vyčleněná kategorie). Vylučujeme i takové neprojektivity, které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

Správné příklady



*… protože doba přenosu **více** závisí na stavu telefonní linky **než** na rychlosti přístroje.* (c101:9)

*s informací o **větším** počtu znaků **než** 16* (c103:6)

***méně** kvalitní vůz **než** jiné západní vozy* (c106:44)

*přibližně **stejně** drahá **jako** ropa* (c109:35)

> Slovo *jako* tady není věšeno jako u doplňků, ale jako podřadící spojka s funkcí AuxC. Morfologickou značku má „J,".
>
> AR: drahá (stejně (přibližně, jako (ropa)))

stejnou** cenovou úroveň **s našimi (c109:39)

***podobné** problémy **jako** u hotelů* (ca11:30)

> AR: problémy (podobné (jako (u (hotelů))))
>
> TR: problém (podobný (&Emp; (on, hotel)))
>
> Konstrukce tedy zůstává neprojektivní i na tektogramatické rovině!

*tak **konkurenční** prostředí **jako** v počítačové branži* (ca16:21)

***jiné** důvody **než** ekonomické* (c214:43)

*Jedná se o **nejdelší** recesi **v moderní historii země**.* (c138:17)

*Obchodníci získají zboží **přibližně** za **stejné** ceny **jako** u výrobců.* (c203:45)

> Kombinace předložkové a srovnávací neprojektivity v jedné frázi.

*… se pozná **spíš** například na tom, jak kdo zachází s nářadím, **než** na výučním listu.* (c129:30)

*Takový podnik působí dojmem **spíš** jakéhosi vetešnictví, **než** solidního obchodu s uměleckými předměty.* (ca10:10)

> Slovo *spíš* visí přímo na slovese, proto jsou dvě neprojektivity zaklesnuté do sebe. Kdyby viselo na *vetešnictví*, což bych považoval za moudré, byla by neprojektivita jenom jedna.

Nechtěné příklady (je to adjektivní či adverbiální fráze, ale nemá srovnávací povahu)

krátká** doba **na hlubší analýzu

*množství **prodaných** lístků **prostřednictvím předprodeje***

*nemá **s jeho hodnotou** nic **společného***

*počet reálně **odpracovaných** hodin **v kalendářním roce***

***donedávna** jeden **společný** státní podnik* (*donedávna* visí na *společný*)

*Jeho ražba byla **zahájená** v roce 1326 **z rozhodnutí** krále Róberta.*

*… redukce **docházejících** formátů **tak**, **aby**…*
Není to spíš chyba anotace? Nemělo by *tak* i *aby* viset na *redukce*?

*Trvanlivost dokumentů je **prakticky** při správném uložení **neomezená**.*
To není srovnání, ale rematizátor, navíc rozdělující skupina je vložena dle mého soudu nevhodně, klidně by mohla být jinde.

*Budu **povinen** za uvedených okolností, tedy …, **stát se plátcem DPH**.*

*Pozdě **podaný** odpor **(**po uplynutí 15 dnů počítaných ode dne doručení) předseda senátu zamítne.* (c135:1)

> Apozice mezi *pozdě* a celým obsahem závorek je řízena levou závorkou, ta visí na *podaný*, tj. na adjektivu, od kterého je oddělena slovem *odpor*. Se srovnáváním to celé nemá nic společného.
>
> Poznámka: Tentýž příklad se objeví i u nově vyčleněné skupiny neprojektivních apozic, a to kvůli druhé neprojektivitě, kterou obsahuje: spojení slova *pozdě* s kořenem apozice (závorkou).

Vyjádření množství

Podobně jako u B3, i tady vzniká podskupina neprojektivit, které vyjadřují množství a silně připomínají číslovkové skupiny A3 nebo B4. Tady jde nejčastěji o příslovce míry, která se chovají podobně jako neurčité číslovky: *hodně* (42 výskytů[3]), *dost* (17), *málo-3* (13), *moc-3* (3), *nazbyt* (2), *minimálně* (2), *příliš* (2), *poskrovnu* (1), *tolik-3* (1), *stejně* (1).

Spokojených zákazníků je dost. (c210:17)

[3] Počítal jsem takové výskyty, kde závislý uzel neprojektivní hrany byl v genitivu. Ani u příslovce *stejně* zde tedy nebyla neprojektivně zavěšená srovnávací fráze, ale počítaný předmět: *Je možné, že jich bude zhruba stejně.* (v112:46).

Otevřených fondů není na peněžním trhu hodně. (c234:1)

Celkem jsem nalezl 139 neprojektivit kategorie B5, jejichž závislý uzel byl v genitivu. Jen část z nich vyjadřovala množství (viz počty výskytů uvedené výše). Všechny jsou zahrnuty i v celkovém počtu neprojektivit typu B5, uvedeném níže.

Chybné příklady (patří do jiné kategorie)

Pro posouzení *budou* ***rozhodující*** *především majetkové poměry.*
Složený jmenný přísudek, patří někam jinam.

Zájmena

Zatím pracovní skupina neprojektivit, které se nechytly výše a jsou řízené zájmenem. Cílem je rozdělit je na srovnávací neprojektivity (B5) a neprojektivity jmenných skupin (B3).

Celkem je jich momentálně 97.

Telefonní linky nemají ***takovou*** *kvalitu* ***jako*** *v laboratoři.* (c101:15)

o ***to*** *více,* ***že***

v ***takovém*** *rozsahu,* ***že*** (l229:10)

takovou *dynamiku,* ***jako***

takové *vlny násilí, jaká* ***zachvátila*** *sousední Rwandu* (la12:26)

tím *méně peněz, čím déle* ***ležel*** (ll27:41)

tím *spíše,* ***že*** (ls10:14)

takovou *autoritu,* ***aby*** (ls29:7)

v ***takové*** *výši, jak to smlouva* ***ukládá***

týmž *závěrům,* ***jako*** (m227:10)

témže *směru,* ***jako*** (v145:45)

týchž *problémech* ***jako*** (v146:21)

tatáž *kvalifikace* ***jako*** (v146:50)

téže *polorovině* ***jako*** (v149:52)

PROTIPŘÍKLADY – normální jmenné skupiny

všichni … z, někteří … z, žádný … z

něco … o nás

nic nepřevážel, co by nebylo uvedeno…

to je všechno otázka zkušeností

je toho poměrně dost, s čím…

které může být, byť oděné do…

zdaleka neobracejí všichni

ten je spolu s…

Celkem z neprojektivit typu B5:

- 352 mělo v závislém uzlu lemma *než-2* nebo *nežli*.
- 175 mělo v závislém uzlu lemma *jako*. Typy *takový-jako*, *stejně-jako*, *stejný-jako*, *podobný-jako*, *tak-jako*, *obdobný-jako*, *týž-jako*, *tentýž-jako*… Teoreticky však

libovolné přídavné jméno: *tak **konkurenční** prostředí **jako** v počítačové branži* (ca16:21).

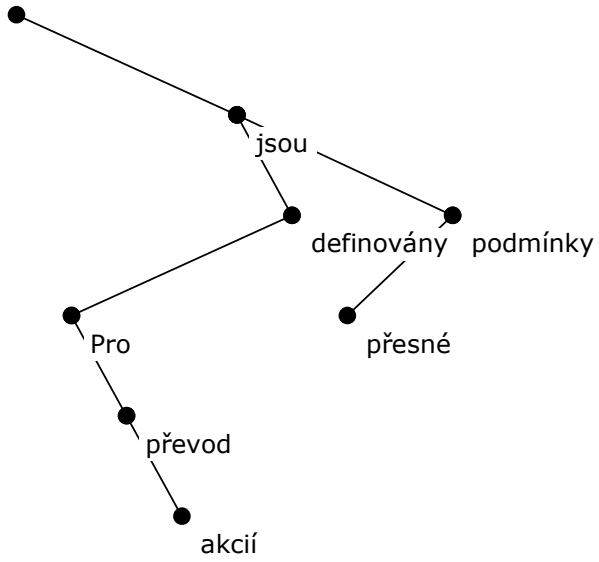
- 84 mělo v řídícím uzlu superlativ. Některé z nich mají srovnávací charakter (*nejdelší recesi v historii* c138:17; *největší výrobce na světě* cb03:2; *nejvhodnější barva pro ženy* l214:36; *nejúspěšnější expozicí za několik let* lb30:2; *nejvyšší průměrné mzdy ze všech odvětví* le02:8). Jiné mají množstevní charakter (*nejvíc je právě oprav podpatků* cb26:33). I tady se najde šum (***potenciálně** jeden z **nejbohatších** latinskoamerických států* le33:27).
- 7 mělo v řídícím uzlu lemma *takový*, ale v závislém nebylo *jako*. *poznatky v **takovém** rozsahu, **že**…* (l229:10). ***takové** vlny násilí, jaká **zachvátila** sousední Rwandu* (la12:26). ***takovou** autoritu, **aby**…* (ls29:7).
- 6 mělo v řídícím uzlu lemma *ten* a řídící uzel závisel na uzlu s komparativem (typy *tím více, čím* a *o to více, že*).
- 8 mělo v řídícím uzlu lemma *stejný* nebo *stejně*, ale v závislém nebylo *jako*: *na **stejnou** cenovou úroveň **s** našimi* (c109:39); případně vedlejší věta: ***stejně** rychle, jak naděje **svitla**, tak rychle pohasla* (ml43:48). Může jít však i o množstevní neprojektivitu (*že **jich** bude zhruba **stejně*** v112:46), nebo prostě o „obyčejné" neprojektivity (***stejné** podmínky **pro všechny*** vb43:24).
- 77 mělo v řídícím uzlu komparativ, ale v závislém nebylo *než-2*. Obvykle nejde o srovnání. Velmi často jde o množstevní konstrukce, podobné číslovkovým neprojektivitám (*Poněkud **více** je **jich** na inženýrském stupni.* ca25:33). Jsou tam však i srovnávací konstrukce typu *o kolik* (*Kdo ví, **o** kolik jsme **chudší**.* cb18:31). Raritou je konstrukce ***vyšší** procento kuřáků **ve srovnání s** vyspělými zeměmi* (le26:1).
- 542 mělo v řídícím uzlu přídavné jméno, zájmeno nebo příslovce, ale nespadá do žádné z výše uvedených kategorií. Pravděpodobně nejde o srovnání.

B5a: neprojektivní rozvití příčestí trpných

Za adjektivní frázi lze do jisté míry považovat i podstrom, v jehož kořeni je příčestí trpné. Příčestí je tvar slovesa a má značku začínající na Vs, neprojektivity tohoto druhu tedy nezachytí filtr skupiny B5. Je to vlastně dobře, protože srovnávací charakter je u příčestí trpných velmi nepravděpodobný (nelze je stupňovat).

Do kategorie B5a jsme zahrnuli všechny neprojektivity, jejichž řídící uzel má značku začínající na Vs a závislý uzel není vztažné slovo (viz B6). Vylučujeme však takové neprojektivity, které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

Neprojektivit typu B5a se v datech našlo 241.



Pro převod *akcií jsou* ***definovány*** *přesné podmínky.* (c121:7)

Nejčastěji bude příčestí trpné součástí přísudku jmenného se sponou.

Občanství ***mi*** *zůstalo* ***zachováno****.* (c204:39)

Konstrukce je podobná, ale sloveso *zůstat* se nepovažuje za sponu.

Na požádání *může být* ***zaslán*** *seznam firem.* (c128:7)

Patří do kategorie A1-modal, ale nedostala se do ní, protože na modálním slovese nevisí infinitiv, ale příčestí trpné.

…, kterému ***soudu*** *je* ***určen*** (c134:12)

Dosud nepodchycená verze B6!

Na zbytek částky musí být složeny předem peníze v bance. (c201:34)

Dvě neprojektivity a obě zmizí! Ta druhá se vymyká dosavadním představám o neprojektivitách s modálním slovesem, protože podmět modálního slovesa (*peníze*) odděluje předmět od významového slovesa (*složeny-v bance*).

B6: vztažné a tázací věty

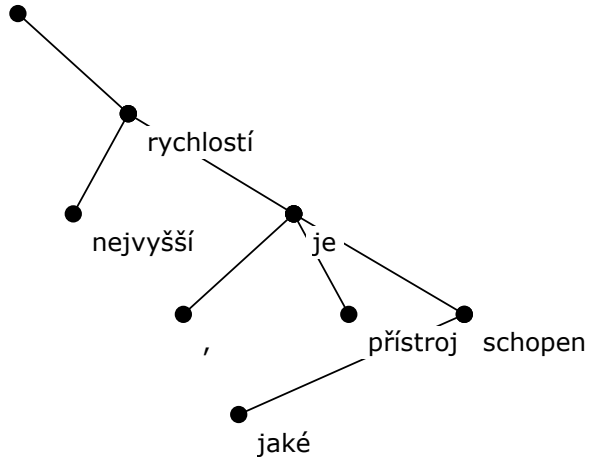
Vztažná (wh) slova jsou oddělena od svého řídícího členu, typicky slovesem. Vztažné slovo se pozná podle lemmatu. Do B6 patří neprojektivity, jejichž závislý uzel má jedno z následujících lemmat:

kdo, co-1, který, jaký, čí, jenž, což-1, kde, odkud, kudy, kam, kdy, jak-3, kolik, kolikátý, kolikerý, kolikrát, nakolik…

Do kategorie B6 nepatří neprojektivně zavěšená vztažná slova na infinitivu pod modálním slovesem, ta patří do kategorie A1-modal. Vylučujeme i takové neprojektivity, které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

Pozor, vztažné slovo může být také schováno pod předložkou, pak buď vzniknou dvě neprojektivity (jedna od předložky k podstatnému jménu, druhá zpět ke vztažnému přívlastku: *Nevím, z jakého odjel nádraží.*), nebo je neprojektivně zavěšená předložka (*Nevím, o kolik vstupenek má zájem.*).

Neprojektivit typu B6 se v datech našlo 136. Nejsou v tom zahrnuty neprojektivity patřící současně do kategorie B7 — průnik těchto kategorií je níže zpracován samostatně.



*nejvyšší rychlostí, **jaké** je přístroj **schopen*** (c101:13)

jaké** z toho plyne **poučení (c112:12)

*Potřeboval bych vědět, **jaký** mám zvolit **postup**.* (ca23:45)

Zájmeno *jaký* se na realizaci neprojektivity typu B6 podílí velmi často.

***Co** je to **platné**?* (c216:50)

*Grafy ukazují, **nakolik** byla tato snaha **úspěšná**.* (ca28:16)

*porušování lidských práv, **jehož** byl denně **svědkem*** (l113:39)

*smrt družky, **kterou** jsem měl **rád*** (lb23:34)

*Dosáhnu alespoň toho, **čeho** jsem **schopen**.* (lb23:62)

*zahraničního partnera, **jemuž** by mělo být **odprodáno** 27 procent akcií* (c235:50)

Nápadně to připomíná neprojektivity s modálními slovesy a infinitivy, až na to, že tady je ještě navíc podmiňovací způsob. Potřebuje dořešit!

Zatím stále neumíme zachytit hlouběji zapuštěná vztažná slova:

o jaký druh práce má zájem

v jak velké firmě chce pracovat

Průnik B6 a B7: vztažná slova závislá na infinitivu

Patří sem neprojektivity, které současně splňují definice B6 i B7. V datech se takových neprojektivit našlo 211.

*záruky, **které** lze v konkurzu **zpeněžit*** (c104:13)

***Co** zamýšlí vaše společnost **udělat** s 15 procenty akcií Tatry?* (c108:40)

*manažerů, **které** se podařilo **získat*** (c139:40)

*několik věcí, **které** by bylo žádoucí **změnit*** (c208:22)

což** je nutné **chápat (c224:41)

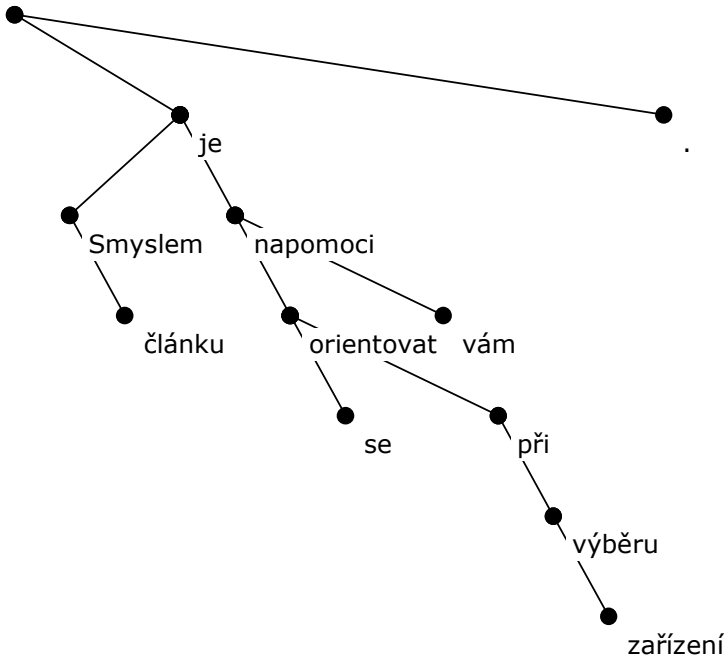
B7: infinitivy

Obecná charakteristika: řídící uzel neprojektivity je sloveso v infinitivu. V díře je něco, na čem tento infinitiv závisí. Nejčastěji je to asi modální sloveso, takové případy ale ve skutečnosti patří do A1. Modální sloveso se totiž (na rozdíl od sloves kvazimodálních) považuje za nesvéprávné funkční slovo a na tektogramatické rovině bude zahubeno.

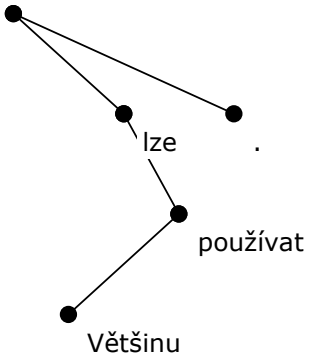
Dochází tak k nepříliš pochopitelnému jevu, že totiž s větou *To nemůžu udělat.* se zachází jinak než s větou *To nezvládnu udělat.* nebo *To nemám možnost udělat.*

Vylučujeme odsud neprojektivity, které zachytilo pravidlo o vztažných slovech (B6) nebo které splňují podmínky příslušnosti k B8 (vysvětlení viz B8).

Neprojektivit typu B7 se v datech našlo 1988. Nejčastějšími rodiči řídícího uzlu neprojektivity jsou uzly s následujícími lemmaty: *lze*, *být*, *začít-1*, *a-1* (88), *začínat* (75), *hodlat* (72), *podařit* (69), *nechat* (53), *snažit* (41), *schopný* (38), *jít* (29), *odmítnout* (27), *stačit* (26), *přestat* (16), *pokusit* (15), *potřebovat* (15), *odmítat* (14), *ochotný* (14), *ale* (13), *povinný* (12), *mínit* (12), *přijít* (12), *nechávat* (11), *pokoušet* (10)…



Smyslem článku je napomoci ***orientovat se*** *vám* ***při výběru*** *zařízení.* (c101:11)



Většinu *lze* ***používat****.* (c101:41)

počtu dlužníků, ***na něž*** *hodlají věřitelé* ***uvalit*** *žalobu* (c103:43)

Rizikové ***úvěry*** *„stíhají" **krýt** zvyšováním rezerv.* (c104:36)

… kdy ***si zajímavé klienty*** *začnou* ***zvát*** *do své pražské pobočky.* (c105:7)

*… že tak velké **množství** trh zřejmě není schopen **absorbovat**.* (c107:25)

*Váš **dotaz** je připraven **zodpovědět** spolupracovník Profitu.*

***Na mobilizaci** je **proto** třeba **pohlížet** jako na opatření výjimečné.*

V tomto případě není *proto* nad celou klauzí, ale visí na *pohlížet*.

***Za jeden forint** již není možné v současnosti **koupit** prakticky nic.*

*Má-li zákazník zájem, může **si** teď už **koupit** Tatru s motorem Deutz.* (c106:48)

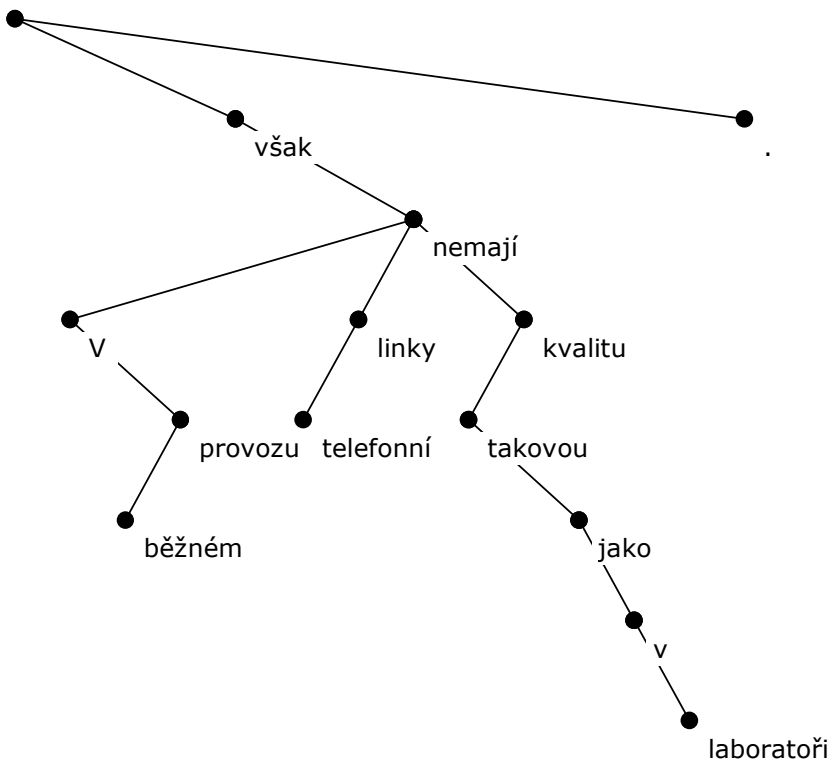
Nedostalo se do A1-modal, protože v díře není přímo modální sloveso *může*, ale jeho potomek *teď už*. Zkontrolovat, co se s takovou větou děje na tektogramatické rovině!

*Společnost začala **poskytovat** v první třídě **místa ke spánku**.*

Zajímavé tím, že v díře neleží samo fázové sloveso, ale jen předložková fráze, která ho rozvíjí. Neprojektivita také nevede zprava doleva, jak je u infinitivů obvyklé. Je ovšem diskutabilní, zda má předložková skupina viset na *začala*, i s ohledem na neprojektivitu by jí spíše slušelo zavěšení pod *poskytovat*.

B8: částice na druhé pozici, hlavy jednočlenných koordinací (však, ale, ovšem, proto, tedy…)

Pracovní hypotéza: Většina neprojektivit tohoto druhu se pozná podle toho, že v díře je jediné slovo s afunem Coord.



V běžném provozu <u>však</u> telefonní linky nemají takovou kvalitu. (c101:14)

S tím <u>ale</u> konstruktéři počítali. (c101:15)

(pod který <u>ovšem</u> Auto Bohemia nespadá) (c126:4)

Ve většině případů má „koordinace" řízená slovem v díře jen jediného člena a pomyslný druhý člen koordinace se nachází někde v předcházejících větách. Ukázalo se však, že podobné neprojektivity mohou být způsobeny i opravdovými koordinacemi, jako v následujícím příkladě:

*Jde o subjektivní rozhodnutí věřitele, a **je** proto **logický postoj** bank, které zvažují, zda se jim bankrot vyplatí nebo ne.* (c103:38)

> Spojka *proto* řídí koordinaci sloves *jde* a *je*. Současně způsobuje neprojektivitu závislostí *je-logický* a *je-postoj*.

U většiny ostatních typů třídy B do definice přidáváme podmínku, aby současně nesplňovaly podmínku příslušnosti k B8. Pokud totiž v díře leží *však* nebo jiné podobné slovo, je skoro vždy příčinou neprojektivity ono a ne faktory, se kterými počítá dotyčná konkurenční kategorie. Pro ilustraci uvádím několik příkladů neprojektivit, které současně s B8 splňovaly podmínky jiných kategorií, ale byly nakonec zařazeny pouze do B8.

Neprojektivit typu B8 bylo v datech nalezeno 3706. Nejčastěji byla zaznamenána tato „děrová" slova: *však* (2526), *ale* (571), *proto* (306), *ovšem* (162), *tedy* (58)…

***Na tom** však **vinu** nemám.* (cb23:38)

> B8+B2-3 (přívlastek). V tomto případě je navíc přívlastek označen jako AtrAdv, tj. podle mínění anotátora by stejně dobře mohl být i příslovečným určením a viset na slovese. Neprojektivita typu B8 by tím však nezmizela.

***Z tvrdé hry soupeřů** však **strach** nemá.* (l106:40)

> B8+B2-3.

***samotná** však **měla** pohledávky za 2,6 miliardy korun* (c104:32)

> B8+B3. Příslušnost k B3 je navíc podezřelá. Jde o doplněk zavěšený na slovese a u nich si pořád nejsem jistý, zda je neprojektivita způsobena tím, že jde o doplněk, nebo jinými faktory. V průniku kategorií B8 a B3 se AtvV (oproti Atv) vyskytuje mnohem častěji než v samotném B3.

*Jako **jedna** z mála světových firem <u>však</u> právě **Rover** nemá oficiální zastoupení.* (lq04:15)

> B8+B3 s normálním doplňkem visícím na podmětu (Atv).

***Kolik** však **z těchto případů** bylo vskutku havarijních?* (l101:42)

> Ve větě jsou dvě neprojektivity, *kolik-z případů* a *bylo-kolik*. Ta první splňuje B8 i B4.

*až do výše 40 %, **maximálně** však **60000** DEM* (va29:37)

> B8+B4, tady je navíc B4 „podivné" (asi nesplňuje to, co jsme v B4 zamýšleli?).

*méně křehké, **ne** však **zralé*** (ce05:37)

> B8+B5

***Podle posouzení** nezávislého auditu byly <u>však</u> velice špatně **strukturovány**.* (c104:33)

> B8+B5a

***Kolik** lidí si však **uvědomí**, že…* (c113:41)

> B8+B6

*Neopomněl **se** však **otřít** o pana Navrátila.* (c137:47)

> B8+B7

B8a: PREC

Slova jako *však*, která způsobují neprojektivity typu B8, dostávají na tektogramatické rovině funktor PREC. Máme seznam takových slov, která tento funktor mohou dostat (i když pochopitelně ne všechna z nich ho dostanou ve všech kontextech, ve kterých se mohou vyskytnout).

A skutečně, když jsem hledal neprojektivity, ve kterých je jediným členem díry slovo ze seznamu PREC, našel jsem k výše uvedeným neprojektivitám typu B8 ještě 65 dalších.

C1: frazémy typu CPHR

CPHR je funktor, kterým se dotyčné frazémy označují na tektogramatické rovině. Ještě tam existují frazémy typu DPHR, které odpovídají naší kategorii B2. U DPHR se počítá s tím, že v budoucnosti budou na tektogramatické rovině reprezentovány jediným uzlem. U CPHR je něco takového těžko představitelné, protože obsahují valenční podstatná jména (např. *zájem o něco*). Typický CPHR frazém obsahuje sloveso a jedno další slovo, často podstatné jméno. Sloveso má na tektogramatické rovině normální funktor (třeba PRED), slovo na něm závislé má funktor CPHR. Pokud frazém způsobuje neprojektivitu, pak je to tím, že se skládá z více uzlů. Jeho rozvití visí na závislém uzlu (a nejde to udělat jinak, protože jde o valenční vazbu), ale je od něj odděleno dírou, ve které věží hlava frazému – sloveso.

V PDT bylo nalezeno celkem 139 neprojektivit typu C1.

Příklad v našem článku *… že ho je třeba přesvědčit …* sice obsahuje CPHR frazém *je třeba*, ale obsahuje také infinitiv, takže u mne zatím skončí v kategorii B7. Vhodnější příklad by byl následující:

*Je naprosto korektní, aby byly **na osoby**, které slouží v armádě, kladeny vyšší **nároky** než na civilisty.* (lq14:9)

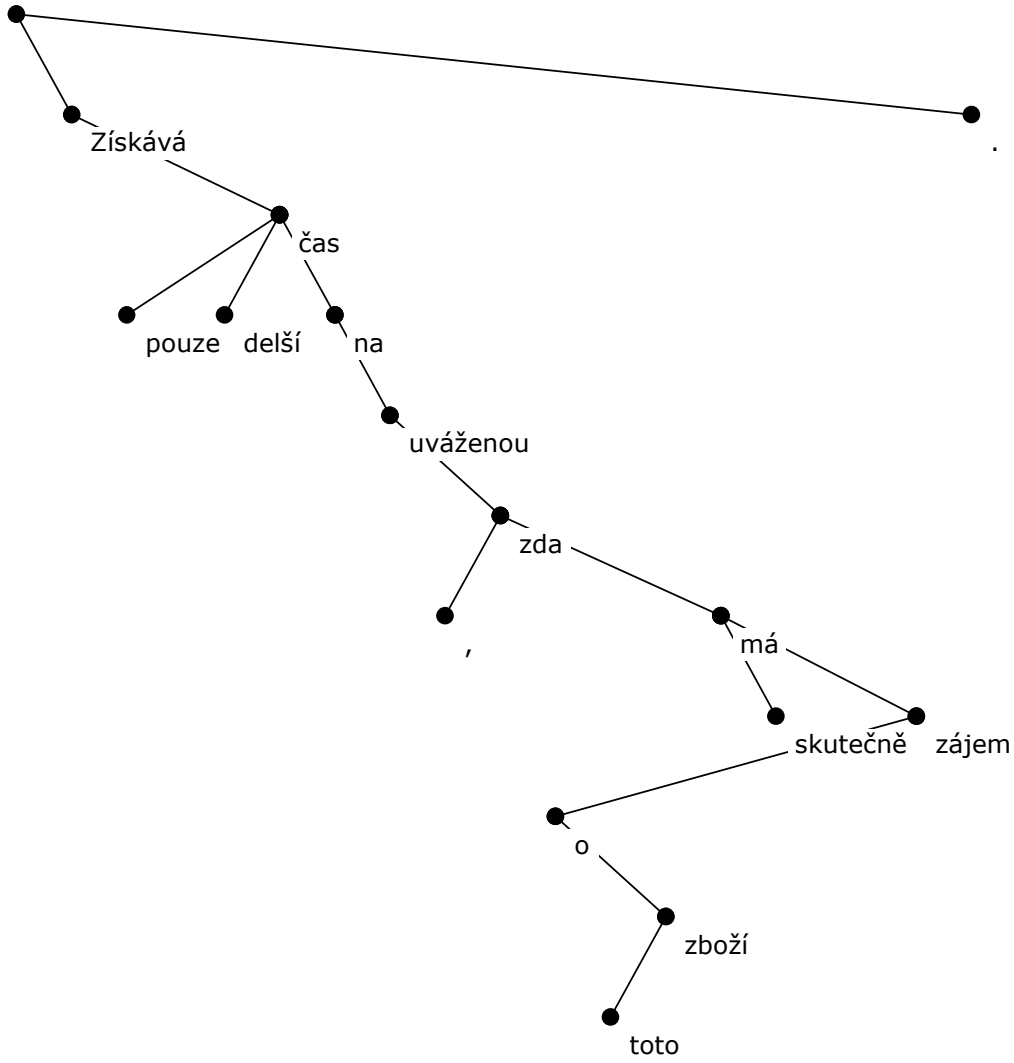
> Frazém typu CPHR *klást nároky na* je uspořádán tak, že závislost *na osoby* na *nároky* vede přes *kladeny*.

Máme k dispozici seznam CPHR frazémů, který dostávají jako pomůcku anotátoři na tektogramatické rovině. I když je tento seznam v principu otevřený, je to jediné vodítko, podle kterého můžeme frazémy strojově poznat. Bez něj by se nám tato kategorie rozpustila v ostatních kategoriích.

Při hledání neprojektivit typu C1 požadujeme, aby sloveso ze složeného přísudku bylo v díře. Jinak by sem spadly i neprojektivity, které sice obsahují složené přísudky, ale nejsou způsobeny jejich složeností:

Podal jsem žalobu u soudu, kterou jsem si sám napsal. (c134:7)

> Složený predikát je *podat žalobu*.



Další příklady CPHR neprojektivit:

Získává pouze další čas na uváženou, zda o toto zboží má skutečně zájem. (c227:16)

Velký zájem má o spolupráci v České republice. (cc02:31)

Určitý šperk vyrábějí, dokud o něj mají zákazníci zájem. (ce21:30)

Řekněme rovnou, že té rybě je úplně jedno, na jak drahé výbavě bude ulovena. (ce25:4)

X-Apozice

Skupinou zatím nezařazenou do jedné ze tříd A, B, C jsou apozice, jejichž první člen je normálně zapuštěný do věty, druhý je pak nějaký výčet nebo popis za dvojtečkou či závorkou na konci věty. Neprojektivní je závislost prvního členu na kořeni apozice (kterým je ona dvojtečka či závorka). Neprojektivních apozic se našlo 130.

*Jedna **věc** je očividná**:** Žatecko je výrazným agrárním regionem.* (c105:2)

***Později** podaný odpor **(**po uplynutí 15 dnů) předseda senátu zamítne.* (c135:1)

*Uvádějí **čistou** pracovní dobu, **tj**. bez polední přestávky.* (cb07:43)

AR: dobu (pracovní, tj (čistou, ,, ., bez (přestávky (polední))))

TR: doba (pracovní, tj (čistý, přestávka (polední)))

Konstrukce zůstává neprojektivní i na tektogramatické rovině!

*Cíl knihy vyjádřil zcela **jasně** již v předmluvě**:** poznání japonského managementu má mnohé výhody…* (cb15:4)

*Každý den s ní musíte alespoň **dvakrát** manipulovat **–** při ranním vytažení a večerním stažení.* (cc24:6)

AR: manipulovat (den (každý), – (dvakrát (alespoň), při (a (vytažení, stažení))))

TR: manipulovat (&Hyphen; (alespoň, dvakrát, a (vytažení, stažení)))

Konstrukce zůstává neprojektivní i na tektogramatické rovině!

X-Ostatní

Odpadkový koš na neprojektivity, které prošly předcházejícím sítem a nespadly do žádné kategorie. Zatím jich je 281. Ideální by bylo rozšířit klasifikaci tak, aby v kategorii „jiné" uvízlo nanejvýš několik desítek stromů, které by odpovídaly anotačním či gramatickým chybám.

Další skupinou budou rematizátory použité jinde než u předložkových skupin (A2):

*… se **také** používá **xeroxový papír** …*

Rematizovaným předložkovým skupinám typu A2 se podobá i následující příklad, ale nemohl do nich být zařazen, protože slovo *jako* má morfologickou značku začínající na J a afun AuxC. Současně má tento příklad společné rysy s doplňky, doplněk to však není (*kopírky* má afun ExD).

*Možnost využití **i** jako **kopírky**.*

Podobně jako apozice mohou neprojektivitu způsobovat vsuvky v díře (viz příklady výše). Tady však vsuvka sama neprojektivně visí, v díře je dvojtečka:

***Rozměry**: **(mm)** 277×180×57*

Závěrečný přehled

Následující tabulka shrnuje jednotlivé typy neprojektivit seřazené sestupně podle počtu výskytů.

Typ	Četnost
A2-předložky	6124
B8-půlkoordinace	3706
A1-modal	3233
B7-infinitiv	1988
A1-li	1089
A4-AuxK	977
B3-doplněk	947
B1-koordinace	835
A5-AuxGXYZ	609
B2-3	572
B5-ostatní	542
A1-složené-spojky	407
A1-AuxV	402
B5-než	352
jiné	281
B4-číslovky	265
B5a-příčestí	241
B6-vztažná+B7-infinitiv	211
B5-jako	175
A3-číslovky	140
C-CPHR	139
B6-vztažná	136
X2-Apos	130
B5-superlativ	84
B8a-PREC	65
B2-DPHR	20
B5-stejný	8
B5-takový	7
B5-tím-více	6
CELKEM	**23691**

Odkazy

Byly publikovány různé definice projektivity, řada z nich je navzájem ekvivalentních. Pojem projektivity údajně zavedl (Lecerf, 1960). Často se cituje (Marcus, 1965); jeho definice údajně koresponduje s definicí (Robinson, 1970). Související pojem má být také *adjacency* (Hudson, 1990). (Kahane et al., 1998) definuje pojem *pseudoprojektivity*.

(Holan et al., 1998) definuje pojem díry a využívá ho k měření složitosti věty z hlediska slovosledu. Formuluje některé hypotézy o maximální složitosti věty v češtině a uvádí některé (umělé) příklady.

Přehled neprojektivních konstrukcí v češtině poprvé podala (Uhlířová, 1972), ovšem ještě bez podkladů z korpusu. Petkevič velmi podrobně rozebírá neprojektivity dle typu zainteresovaných uzlů a uvádí řadu příkladů, někdy příliš umělých. (Holan, 2003) předkládá některé statistiky ohledně neprojektivit v PDT, vesměs jde však pouze o kumulativní čísla ohledně výskytu konkrétních morfologických značek v neprojektivitách apod. Statistiky neprojektivit v PDT se poprvé objevily v (Hajičová et al., 2004), tam však bylo přednostním cílem zkoumání důvodů, proč se neprojektivity objevily a zda se mohou vynořit i na *ideální* tektogramatické rovině (tedy na TR v takovém tvaru, v jakém bychom ji chtěli mít, kdyby lidské, finanční a jiné zdroje byly neomezené). Technická zpráva, kterou držíte v ruce, z tohoto článku vychází, přebírá a rozšiřuje jeho klasifikaci.

Literatura

Jan Hajič, Barbora Vidová Hladká, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas (2001): *Prague Dependency Treebank 1.0 CD-ROM.* Catalog # LDC2001T10, ISBN 1-58563-212-0. Linguistic Data Consortium, Philadelphia, Pennsylvania.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, Daniel Zeman (2004): *Issues of Projectivity in the Prague Dependency Treebank.* In: Prague Bulletin of Mathematical Linguistics, vol. 81, pp. 5–22. ISSN 0032-6585. Univerzita Karlova, Praha, Czechia.

Eva Hajičová, Jarmila Panevová, Petr Sgall (2000): *A Manual for the Tectogrammatical Tagging of the Prague Dependency Treebank.* Technical Report TR-2000-09. Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha, Czechia.

Tomáš Holan (2003): *K syntaktické analýze českých (!) vět.* In: David Obdržálek, Jana Tesková (eds.): MIS2003 Josefův Důl, sborník semináře, pp. 66–74. Matfyzpress, Praha, Czechia.

Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek (1998): *Two Useful Measures of Word-Order Complexity.* In: A. Polguere, S. Kahane (eds.): Proceedings of the COLING'98 Workshop on Processing of Dependency-Based Grammars, pp. 21–28. Université de Montréal, Montréal, Québec.

Richard Hudson (1990): *English Word Grammar.* Basil Blackwell, Oxford, Britain.

Sylvain Kahane, Alexis Nasr, Owen Rambow (1998): *Pseudo-projectivity: a polynomially parsable non-projective dependency grammar.* In: Proceedings of the 17th International Conference on Computational Linguistics (COLING 98), pp. 646–652. Université de Montréal, Montréal, Québec.

Vladislav Kuboň, Tomáš Holan, Karel Oliva, Martin Plátek (2001): *Word-Order Relaxations & Restrictions within a Dependency Grammar.* In: Proceedings of International Workshop on Parsing Technologies (IWPT), pp. 237–240. 清华大学出版社, 北京, China.

Yves Lecerf (1960): *Programme des conflits, modèle des conflits.* In: Bulletin bimestriel de l'ATALA, 4, 5.

Solomon Marcus (1965): *Sur la notion de projectivité.* In: Zeitschrift für mathematische Logik und Grundlagen der Mathematik, vol. 11, pp. 181–192, ISSN 0044-3050. Leipzig, Germany.

Vladimír Petkevič: *Přehled neprojektivních konstrukcí v češtině (nepublikovaný rukopis).* Ústav teoretické a komputační lingvistiky, Filozofická fakulta, Univerzita Karlova, Praha.

Jane J. Robinson (1970): *Dependency structures and transformational rules.* In: Language, vol. 46(2), pp. 259–285. ISSN 0097-8507. Linguistic Society of America, Baltimore, Maryland.

Ludmila Uhlířová (1972): *On the non-projective constructions in Czech.* In: Prague Studies in Mathematical Linguistics, vol. 3, pp. 171–181. Academia, Praha, Czechia.

Kateřina Veselá, Jiří Havelka, Eva Hajičová (2004): *Condition of Projectivity in the Underlying Dependency Structures.* In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004). Université de Genève, Genève, Switzerland.