Oponentský posudek disertační práce Mgr. Jiřího HAVELKY

Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax

Předložená disertační práce se zabývá matematicko-lingvistickým problémem *par excellence*: neprojektivitou v zakořeněných a (úplně) uspořádaných především závislostních stromech, které představují klasickou matematickou strukturu vhodnou k popisu zejména syntaktických struktur přirozených jazyků. Přesněji, práce zkoumá vztahy mezi uspořádáním uzlů, tedy z jazykovědného hlediska slovosledem, a hranami stromů, tedy z jazykovědného hlediska závislostí, stručněji: (ne)projektivitu a její různé podoby a vlastnosti.

Těžištěm práce je tedy matematický popis (ne)projektivity závislostních stromů, lingvistická stránka slouží jen k ověření platnosti matematického výzkumu, je však také důležitá. Jedním z cílů práce je však matematicky zkoumat neprojektivní syntaktické struktury v přirozených jazycích.

Práce má dvě hlavní části: *teoretickou matematickou* a *empirickou jazykovědnou*. V úvodu práce jsou uvedeny hlavní cíle práce, kterými podle autora jsou:

V části teoretické:

- nová definice (ne)projektivity v závislostních stromech
- vlastnosti neprojektivních hran sloužící k popisu neprojektivních stromů obecně. Dále charakterizace závislostních stromů co do (ne)projektivity, planarity/planárnosti a dobré zapuštěnosti.

V části empirické jazykovědné:

- aplikace teoretických výsledků z hlediska řady matematických vlastností závislostních stromů na reálná data v jazykových korpusech 19 různých, často velmi odlišných přirozených jazyků.

Poté autor přehledně popisuje, na co se má čtenář těšit: teoretická část je obsažena v kapitolách 1 až 8, část empirická pak v kapitolách 9 a 10. V závěru práce je uveden rejstřík hlavních pojmů a označení a použitá literatura.

Zaměřme se nyní stručně na jednotlivé kapitoly a jejich hlavní momenty a výsledky.

Kapitola 1

V této úvodní kapitole nazvané *Preliminaries* autor definuje známé základní pojmy jako zakořeněný a závislostní strom, které bude potřebovat v dalších částech práce. Uvádí i svou představu datové reprezentace stromů pro jejich

implementaci v několika algoritmech, jež tvoří součást teoretické partie práce.

Kapitola 2

Tato kapitola je nazvána *Projectivity and basic properties of non-projective edges*. Autor uvádí tři staré ekvivalentní definice (ne)projektivity závislostních stromů podle Solomona Marcuse (definici Harperovu a Haysovu, Lecerfovu a Ihmovu, a konečně Fitialovovu). Dále definuje pojem díry či otvoru (*gap*), známý např. z práce Holanovy et al. a Bodirského et al., jako množinu „neprojektivních uzlů"[1] (tj. uzlů způsobujících neprojektivitu); každá neprojektivní hrana musí mít nějakou strukturně „cizorodou" díru, jež právě způsobuje neprojektivitu. Autor představuje i algoritmus pro vyhledávání děr v závislostních stromech. Tento algoritmus i všechny následující algoritmy také opatřuje dokázaným údajem o jejich časové složitosti.

Kapitola 3

V kapitole třetí nazvané *Projectivity and projective orderings of dependency trees* formuluje autor projektivitu jinak než dosavadní literatura. Činí tak v podobě pěkně dokázané věty (Věta 3.1.1), v níž vychází na rozdíl od klasických výše zmíněných ekvivalentních definic projektivity shrnutých Marcusem pouze ze strukturních vztahů závislosti mezi uzly v závislostním stromě, čili z hran stromového grafu, a na jejich základě vymezuje vzájemné (slovosledné) uspořádání uzlů, tedy jinou, slovoslednou, nestrukturní dimenzi. Ve Větě 3.1.1 se právě dokazuje ekvivalence nově vymezené projektivity s klasickou. Doktorand rovněž definuje tzv. *kanonickou* a *obecnou projektivizaci* zakořeněných stromů na základě lokálního uspořádání keřů (nejjednodušších lokálních podstromů) stromu. Ukazuje, že pro každý zakořeněný strom existuje *jediná obecná projektivizace* a pro každý závislostní strom *jediná kanonická projektivizace*. Rovněž dokazuje ekvivalenci projektivního uspořádání závislostního stromu a částečného uspořádání stromu na základě uspořádání keřů. V závěru kapitoly uvádí podrobný algoritmus obecné projektivizace zakořeněných stromů spolu s jeho lineární časovou složitostí. Algoritmus je odůvodněn, diskuse vyčerpávající.

Kapitola 4

Tato kapitola je nazvána *Level types of non-projective edges* a autor tu dává do souvislosti vlastnosti děr definovaných výše a úrovně uzlů, tj. jejich relativní vzdálenosti od úrovně 0, což je úroveň kořene závislostního stromu. Definuje pojem typu úrovně (*level type*, úrovňový typ) jako celé číslo charakterizující

[1] Můj pojem, nikoli doktorandův.

největší úrovňovou vzdálenost neprojektivní hrany (od úrovně jejího závislého členu) od díry (velmi zjednodušeně řečeno jde o to, zda se při běžném znázornění nachází uzel způsobující neprojektivitu hrany pod touto hranou, na její úrovni nebo nad ní, a navíc v jaké vzdálenosti). Autor si všímá trichotomie definovaných úrovňových typů (kladné číslo, nula, záporné číslo) a správně tvrdí (Věta 4.1.3), že v závislostním stromě implikuje existence neprojektivní hrany nekladného typu existenci jiné neprojektivní hrany nezáporného typu. Pochopení pojmu mi poněkud ztížil levý obrázek v „(b) level type 0" na straně 40. Neměl tam být ve skutečnosti obrázek ze starší verze práce, kterou jsem měl k dispozici? Mimoto bych byl rád, kdyby autor při obhajobě uvedl ve vztah pojem úrovňový typ (*level type*) s trichotomií (*projektivita*, *slabá neprojektivita*, *silná neprojektivita*) ruských emigrantů Dikovského a Modinové. Na základě úrovňových typů a Věty 4.1.3 formuluje autor nutnou a postačující podmínku projektivity závislostních stromů: *projektivní strom neobsahuje neprojektivní hrany nezáporného úrovňového typu*. Proto doktorand dále definuje pojem horní díry (*upper gap*), neboť se stačí omezit jen na „neprojektivní uzly" ležící v závislostním stromě na nižší úrovni, tedy blíže kořeni stromu, to jest v klasickém znázornění (s kořenem nejvýše) naopak výše. Věta 4.1.7 je pak pouhým přeformulováním Věty 4.1.3 pomocí pojmu horní díra. Následuje Algoritmus (4, 5) vyhledávající neprojektivní hrany nezáporného úrovňového typu, a to odspodu, tedy *bottom up*, což je v algoritmu věc zcela zásadní. Algoritmus je to elegantní, neboť rušením uzlů na nižších úrovních stromu, než je úroveň právě zpracovávaná, se obchází nutnost prověřovat podřízenost (tj. reflexivní a tranzitivní uzávěr závislosti). Autor myšlenku algoritmu jako obvykle podrobně objasňuje a vyjadřuje se rovněž k jeho lineární časové složitosti, kterou ideově dokazuje. Algoritmus nevyhledává neprojektivní hrany záporného typu, ani neidentifikuje úrovňový typ (tj. celé číslo) nalezených neprojektivních hran. Na jeho základě však autor ukazuje, jak efektivně lze zjistit, zda neprojektivní hrana je či naopak není záporného úrovňového typu. Dále autor formuluje Algoritmus 6, v němž koncizně spojuje vyhledávání neprojektivních hran nezáporného úrovňového typu s nalezením kanonické projektivizace závislostního stromu a tvrdí, že jeho časová složitost je lineární.

Kapitola 5

Tato kapitola se zabývá neprojektivními hranami jakožto vůbec nejjednoduššího pojmu ve vztahu k planaritě/planárnosti. Uvádí jednoduchou definici planarity a neplanární dvojice hran (Definice 5.1.1) a ukazuje, jak z planárního uspořádaného stromu „vyrobit" neprojektivní závislostní strom volbou jednoho uzlu na hlavní cestě (navazující posloupnosti hran) planárního uspořádaného stromu za kořen závislostního stromu. Na s. 53 uvádí názorný příklad, který je však po mém soudu v jedné své části chybný: závislostní strom

v (b) není, myslím, „vyroben" správně: z kořene vychází 5 hran, což nemá oporu v původním planárním uspořádaném stromě. Prosím, aby – pokud se nemýlím – autor při obhajobě uvedl obrázek na pravou míru. V kapitole 5 formuluje zejména pomocí pojmu neplanární dvojice hran nutnou podmínku projektivity závislostního stromu, totiž *planaritu*. Má-li však vzniklý závislostní strom svůj kořen jako nejlevější či nejpravější uzel, splývá projektivita s planaritou. Důkaz je kratičký a hezký. Dále prostřednictvím horní díry formuluje doktorand rovněž nutnou a postačující podmínku pro charakterizaci neplanární dvojice hran (Věta 5.2.8). Autor též ukazuje, jak charakterizovat neplanaritu pouze na základě vlastností jednotlivých neprojektivních hran. Na s. 56 vymezuje pomocí díry pro danou hranu *h* závislostního stromu tzv. *neplanární množinu hran*. Ukazuje, že je-li tato množina neprázdná, je hrana *h* neprojektivní. Obrácené tvrzení pochopitelně neplatí. Podstatným výsledkem je však následující charakteristika neplanárního závislostního stromu ZS, totiž neprázdnost neplanární množiny hran u nějaké neprojektivní hrany závislostního stromu ZS. Na základě předchozích výsledků je však možné se omezit na takovou neplanární množinu hran dané neprojektivní hrany *h*, jejichž uzly se – zhruba řečeno – nacházejí v horní díře hrany *h* (Věta 5.3.6). V závěru obsažné kapitoly doktorand prověřuje planaritu v Algoritmech 7 a 8, které pro vstupní závislostní strom vyhledávají všechny jeho (horní, v Algoritmu 8) neplanární hrany s časovou kvadratickou složitostí. Nástin tohoto algoritmu je na straně 58, těsně pod ním jsem však nalezl dvě nejasnosti: (a) odkaz na Definici 6.3.2 je patrně nesprávný – nemá to snad být Definice 5.1.1? Vyjasnit, prosím!; (b) test na řádce 3 zabere údajně konstantní čas. Proč, prosím? Snad díky předkompilaci? V odstavci nazvaném *Remark on NP-completeness of multiplanarity* na s. 60 se definuje *planarita* uspořádaného neorientovaného grafu. Myslím, že chybně: je to spíše definice *neplanarity* uspořádaného neorientovaného grafu. Vyjasnit, prosím!

Kapitola 6

V této kapitole se doktorand zabývá neprojektivními hranami ve vztahu k dobré/špatné zapuštěnosti (*well-nestedness/ill-nestedness*) dvou závislostních stromů (zejména jakožto podstromů daného závislostního stromu). Tento vztah charakterizuje jak na základě dvojice neprojektivních hran, tak prostřednictvím jediné neprojektivní hrany. Uvádí i algoritmus prověřující dobrou zapuštěnost jednoho stromu do druhého. Na rozdíl od klasické definice Bodinského et al. (Definice 6.1.1) vychází od dvojice hran ve stromech, nikoli od stromů jako takových. Špatnou zapuštěnost takto charakterizuje pomocí lemmatu 6.2.1, konkrétně pomocí hran jakožto reprezentantů každého z obou stromů. Na základě Věty 6.3.1 formuluje na s. 65 v Důsledku 6.3.3, Větě 6.3.4 a Důsledku 6.3.5 důležitou charakterizaci špatně zapuštěného závislostního stromu jen

pomocí špatně zapuštěné dvojice hran (*ill-nested pair of edges*), jež jsou definovány prostřednictvím děr; hle, díra – jak se stále ukazuje – je klíčovým základním pojmem nesmírně vhodným pro vymezování vlastností neprojektivních stromů! V odst. 6.4 ve Větě 6.4.1 je uvedena důležitá postačující podmínka: nachází-li se v závislostním stromě neprojektivní hrana nekladného úrovňového typu, je špatně zapuštěna. V odst. 6.5 jde doktorand ve svém zjednodušování ještě dále: ukazuje, že dobrou zapuštěnost lze charakterizovat pouze pomocí jednotlivých neprojektivních hran, nikoli dvojic hran, neřku-li stromů! Vymezuje pojem *množiny špatně zapuštěných hran* (*ill-nested set*) pro danou hranu v závislostním stromě (lapidárně řečeno: vyjadřuje po mém soudu *křížení hran*, nechť mě autor případně opraví!) a ve Větě 6.5.4 tvrdí, že hrany tvořící špatně zapuštěnou dvojici hran s danou hranou patří právě do *množiny špatně zapuštěných hran*. Poté na základě definice *množiny horních špatně zapuštěných hran* či *horní množiny špatně zapuštěných hran* (můj český překlad anglického *upper ill-nested set*) a její neprázdnosti charakterizuje špatně zapuštěný závislostní strom, děje se tak ve Větě 6.5.7. Poté jako obvykle algoritmicky testuje dobrou zapuštěnost, a to v odst. 6.6 na s. 68 (na 5. řádce odspodu není patrně správný odkaz na „Theorem 6.5.5", spíš by mělo být „Corollary 6.5.5" nebo „Theorem 6.5.4"), konkrétně v Algoritmu 9 nachází množiny špatně zapuštěných hran pro neprojektivní hrany závislostních stromů a v Algoritmu 10 nachází množiny horních špatně zapuštěných neprojektivních hran závislostních stromů. Oba algoritmy mají kvadratickou časovou složitost.

Kapitola 7

V této kapitole se doktorand zabývá strukturou díry dané neprojektivní hrany z hlediska *intervalů* (spojité posloupnosti intervenujících neprojektivních uzlů) a strukturních *komponent* (jde v podstatě o podstromy v díře) a definuje *stupeň intervalu* (*interval degree*, počet intervalů v díře) a *stupeň komponenty* (*component level*, počet komponent v díře). Dále definuje tzv. *úrovňovou signaturu* (*level signature*) pro danou hranu jako množinu s případně se opakujícími prvky (*multiset*), jež charakterizuje vzdálenost závislého uzlu dané hrany od kořenů podstromů (může jich být víc) v díře. Typ úrovně je pak maximum v tomto multisetu. Po Větě 7.3.2 směřuje autor k jejímu důsledku (Corollary 7.3.3), kde uvádí postačující podmínku pro špatnou zapuštěnost: stačí aby závislostní strom obsahoval neprojektivní hranu s nekladnou úrovní komponenty (*component level*) ve své úrovňové signatuře.

Kapitoly 4 až 7 tvoří ideové jádro výtečné Havelkovy práce. Řekl bych, že celou prací se vine tato myšlenková nit: dosáhnout z minima maximum, tj. s

pomocí těch nejjednodušších pojmů, konkrétně pojmu hrany, uspořádání a úrovně vzdálenosti uzlu od kořene, definovat projektivitu, zjistit konkrétní místa, jádra a zdroje neprojektivity v závislostním stromě, vymezit vztah neprojektivity k planaritě a špatné zapuštěnosti, obecně řečeno: „jít až k pramenům". Autorovi se to velmi dobře podařilo. Algoritmy prověřené na rozsáhlém reálním jazykovém materiálu a vyplývající z teoretických výsledků předvádějí praktickou aplikaci hlubokých teoretických myšlenek.

Kapitola 8

V této kapitole autor na základě literatury sumarizuje kardinality projektivních stromů o *n* uzlech, planárních uspořádných nezakořeněných stromů o *n* uzlech, planárních uspořádaných zakořeněných stromů o *n* uzlech a dobře zapuštěných závislostních stromů o *n* uzlech v přehledné tabulce. Touto kapitolou končí teoretická matematická část práce.

Kapitola 9

Tato kapitola je první ze dvou kapitol empirické lingvistické části. Dočítám se, že autor své algoritmy prověřil na datech korpusu PDT 2.0. Ukazuje se, že Algoritmus 5 vyhledávající neprojektivní hrany s nezápornými úrovňovými typy a mající lineární časovou složitost pracuje s reálnými českými texty efektivně a rychle, neboť reálná data jsou relativně jednoduchá, jak také dokládají čísla.

Kapitola 10

V této kapitole doktorand empiricky ověřuje pojmy a vztahy, jež zpracoval v teoretické části, na 19 jazycích světa. Autor budí velmi potěšitelný dojem, že zpracovaných 19 přirozených jazyků slušně ovládá, zejména (ne)projektivní vlastnosti jejich syntaktických struktur. Jeho záběr je rozsáhlý, zpracoval data ve stromových jazykových korpusech těchto přirozených jazyků, a to od baskičtiny přes arabštinu po turečtinu, přičemž neopomíjí ani jazyky slovanské, germánské a románské. Z vyhodnocení vysvítá metodologická výhodnost doktorandem zvoleného postupu: s čím jednodušším a lokálnějším pojmem se pracuje, tím jemněji lze provádět analýzu syntaktických struktur příslušných jazyků. Je jasné – a autor je si toho dobře vědom –, že číselné údaje týkající se neprojektivity uvedené s nevšední úrovní podrobnosti a s mimořádnou akribií pro 19 jazyků lze jen obtížně vzájemně srovnávat, neboť syntaktické značkování větných struktur v těchto jazycích je velmi odlišné. Na jednotlivé číselné údaje vrhá jasnější světlo třístránková diskuse na s. 121—123 zabývající se vlastnostmi zkoumaných *stromů* a *hran*. Vyplývá z ní několik závěrů:

Ad stromy:

- druhy neprojektivity jsou z lingvistického hlediska velmi různorodé
- planarita a neprojektivita spolu těsně souvisejí
- nejneprojektivnější z hlediska stromů je latina (excerpován M. T. Cicero, C. J. Caesar, P. Vergilius Maro, sv. Jeroným), vysokou stromovou neprojektivitu vykazuje i holandština (překvapivě!)
- špatně zapuštěných stromů je velmi málo, avšak více, než uvádí literatura
- latina, ergativní baskičtina a němčina mají poměrně hodně špatně zapuštěných závislostních stromů; to asi odpovídá pravdě, o vlastnostech baskičtiny však nedokážu náležitě přemýšlet.

Ad hrany:

Autor dokládá, že detekce neprojektivních hran umožňuje nalézat konkrétní příčinu neprojektivity a pořídit jemnou klasifikaci neprojektivit. Dále uvádí, že

- stupně intervalu a stupně komponenty jsou obecně nízké, ačkoli pro baskičtinu, češtinu, angličtinu a švédštinu jsou čísla relativně vysoká
- kladné úrovňové typy jsou důležitým měřítkem neprojektivity a výrazně předčí typy záporné
- baskičtina má nejvíc špatně zapuštěných závislostních stromů (zjištěno na základě hran).

Kapitola končí výstižným závěrem: vlastnosti neprojektivních hran jakožto nejjednoduššího a zcela základního pojmu spolu s úrovněmi uzlů v závislostním stromě jsou zásadní pro jemný popis syntaktické struktury přirozených jazyků. *Vlastnosti hran lokálně modelují globální jevy*: toto klíčové doktorandovo tvrzení se podařilo přesvědčivě dokázat! Ukázalo se dále, že neprojektivity v reálných přirozených jazycích v podstatě se dají optimálně zachytit úrovňovými signaturami neprojektivních hran.

Neobyčejně by mě zajímal podrobný lingvistický rozbor matematicky nevšedně důkladně zkoumaných typů neprojektivit pro různé jazyky, to však nebylo cílem práce. I tak však autor odvedl obrovskou práci, vždyť srovnat 19 jazyků na základě jejich korpusů a vlastních měr je úctyhodný výkon. Klobouk dolů, doktorande!

Rejstřík

Možná by bylo vhodné doplnit ještě slovníček českých ekvivalentů ne úplně

základních anglických výrazů.

Použitá literatura

Seznam použité literatury je rozsáhlý a – pokud mohu soudit – nechybí v něm žádný relevantní text z období od 60. let 20. století dodnes. Je výborné, že autor dokázal pracovat i s nejnovější literaturou: v soupisu jsou zahrnuty i materiály z let 2003—2007, například práce Joachima Nivreho, Kuhlmanna a Möhla a dalších.

Nejasnosti a sémantické/logické nepřesnosti

Byl bych rád, kdyby disertant při obhajobě podrobněji vysvětlil, doplnil či upřesnil následující místa ve své práci (o některých z nich jsem se už zmínil výše, zde podávám celkové shrnutí):

- na s. 15, 8. řádek shora: posledním písmenem má být snad *i*, nikoli *n*.
- na s. 16, 1. řádek zdola má snad být: $i \rightarrow j_2$, nikoli $j_2 \rightarrow i$
- v důkazu Věty 3.1.1 na s. 29 uprostřed stojí ... , $y_m = v$, má však být pravděpodobně $z_m = v$
- na s. 32 dole na řádku 5 zdola, který obsahuje označení posloupnosti obecných projektivizací, nemá být zřejmě G, nýbrž gotické(?) L
- levý obrázek v „(b) level type 0" na s. 40 asi není správný
- na s. 53: závislostní strom v (b) není, myslím, správně „vyroben"
- pod algoritmem na s. 58 jsou dvě nejasnosti: (a) odkaz na Definici 6.3.2 je patrně nesprávný – nemá to být snad Definice 5.1.1? (b) test na řádce 3 zabere údajně konstantní čas. Proč?
- Na s. 60 v odstavci nazvaném *Remark on NP-completeness of multiplanarity* se definuje *planarita* uspořádaného neorientovaného grafu. Asi chybně: je to spíše definice *neplanarity* uspořádaného neorientovaného grafu. Vyjasnit, prosím!
- na s. 68 na 5. řádce odspodu není patrně správný odkaz na „Theorem 6.5.5", spíš by mělo být „Corollary 6.5.5" nebo „Theorem 6.5.4".

Drobné chyby, zejména překlepy

- na s. 10, 4. řádek shora: místo *generel* má být *general*
- na s. 18, začátek 5. odstavce má být *thesis* nikoli *theses*, a slovo *hi* tam asi nemá být
- na s. 127, 2. řádek odspodu: má být nejspíš *Meeting* místo *Meering*

To jsou jen některé překlepy z velmi malého množství, další tu nezmiňuji. Při čtení práce jsem totiž pořizoval pro autora pečlivou korekturu, s níž ho před obhajobou seznámím, aby práci ještě vybrousil před eventuálním knižním vydáním, jež si práce rozhodně zaslouží.

Jazyková úroveň

Práce je psána po mém soudu velmi dobrou angličtinou, řekl bych, ač nejsem rodilý mluvčí angličtiny, že až bezchybnou.

Závěr

Na závěr konstatuji, že přes drobounké nedostatky formálního rázu oponovaná disertační práce více než přesvědčivě dokazuje schopnosti kandidáta myslet navýsost samostatně a vskutku tvůrčím způsobem řešit klasické, nicméně dost nesnadné problémy v oblasti syntaxe přirozených jazyků, zejména jejich matematické aspekty. Mám za to, že žádné další zkoumání matematických vlastností (ne)projektivních konstrukcí v budoucnu nemůže Havelkovu práci pominout, je totiž stěžejní. Jsem prací a jejími výsledky velmi potěšen, ba nadšen, ale na závěr se vyjádřím spíše koženě, jak je předepsáno: **doporučuji, aby doktorandovi Mgr. Jiřímu Havelkovi byl na základě této nadmíru zdařilé práce i ostatních splněných požadavků kladených na jeho doktorské studium udělen titul Ph.D**.

V Praze dne 3. srpna 2007

doc. RNDr. Vladimír Petkevič, CSc.
Ústav teoretické a komputační lingvistiky FFUK