

## Zápis

z obhajoby disertační práce Mgr. Tomáše Jelínka

konané dne 25. června 2012

Téma práce: „Forma a funkce u substantiv v češtině: vztah pádu a syntaktické funkce  
*Na materiálu korpusu současné psané češtiny (SYN2005)*“

přítomni:

prof. PhDr. František Čermák, DrSc., předseda komise

prof. PhDr. Karel Kučera, CSc., člen komise

doc. RNDr. Markéta Lopatková, Ph.D., oponentka

prof. PhDr. Jarmila Panevová, DrSc., členka komise

doc. RNDr. Vladimír Petkevič, CSc., člen komise a školitel doktoranda

prof. PhDr. Oldřich Uličný, DrSc., oponent

RNDr. Milena Hnátková, CSc., zapisovatelka průběhu obhajoby

Ing. Alexandr Rosen, Ph.D.

Předseda komise prof. PhDr. František Čermák zahájil přesně ve 13:00 obhajobu a představil přítomným kandidáta.

Školitel doc. Vladimír Petkevič představil doktoranda a seznámil komisi s průběhem jeho studia a s jeho odborným působením. Poté pronesl vyjádření školitele.

Kandidát Mgr. Tomáš Jelínek seznámil přítomné se svou disertační prací. Představil hlavní cíl své práce, totiž zpracovat z frekvenčního hlediska vztah syntaktických funkcí substantiv a jejich realizace prostými a předložkovými pády. Stručně shrnul výsledky dvou částí své práce: první část popisuje metodu automatické syntaktické anotace korpusu včetně zapojení vlastního programu pro opravu výsledků stochastického parseru. Ve druhé části jsou prezentovány výsledky práce: frekvenční tabulky syntaktických funkcí substantiv ve vztahu k jejich pádu, podrobný rozbor všech relevantních kombinací.

Poté přednesli závěry svých posudků oponenti práce doc. Markéta Lopatková a prof. Oldřich Uličný. Doc. Markéta Lopatková zdůraznila, že jako hlavní přínos práce hodnotí návrh a

implementaci opravného modulu pro automatickou syntaktickou analýzu. Konstatovala, že autor zejména ve svém popisu opravných pravidel dokazuje hluboký vhled do problematiky automatické syntaktické analýzy a do postupů kvantitativní lingvistiky. Vysoce hodnotila zejména hybridní přístup autora – po základním zpracování vstupu statistickým parserem následuje modul sestávající z pravidel, která odhalují chybnou analýzu a případně realizují opravy syntaktické struktury, syntaktických funkcí a morfologických značek. Za velice přínosné pokládá též testy na obsáhlých vzorcích substantiv s jednotlivými formami, mj. také ocenila koncepční přístup k řešené problematice, schopnost samostatné práce s rozsáhlými jazykovými daty a jejich automatickým zpracováním.

Za podstatný nedostatek považuje doc. Markéta Lopatková chybějící zasazení do kontextu současných bohemistických prací a vyrovnání s teoretickými přístupy k dané problematice. Doc. Lopatková také vznesla několik dotazů: zajímala se o způsob využití existujících jazykových zdrojů, o pořadí opravných pravidel při jejich aplikaci na text a také o úspěšnost opravných pravidel při jejich aplikaci na optimálně nastavený parser.

Svůj posudek shrnula tak, že práce sice dokazuje hluboký vhled do problematiky (automatické) syntaktické analýzy a do postupů kvantitativní lingvistiky, ale chybějící zasazení do kontextu bohemistických prací považuje za závažný nedostatek, rozhodnutí ponechává na komisi.

Prof. Oldřich Uličný rovněž velmi ocenil to, že disertant dokázal kvalitně skloubit lingvistickou práci s prací programátorskou, statistickou a korpusovou. Konstatoval, že práce využívá existujících i vlastních počítačových prostředků k analýze jazykových struktur češtiny zachycených v korpusu a přináší řadu nových poznatků i analytických postupů.

Prof. Uličného zaujala zvláště zjištění o poměru realizace jmenné části verbonominálního predikátu nominativem a instrumentálem, které ukazuje, že instrumentál (aspoň v odborné literatuře) není na ústupu. Byl také překvapen, že disertační práce uvádí výskyty prostého akuzativu ve funkci atributu a žádal o příklady.

Prof. Oldřich Uličný upozornil na terminologický problém v používání termínu *forma* místo vhodnějšího *výraz* a také na nevhodné použití termínu *pád* místo *pádový tvar* nebo *deklinační tvar*.

Dále položil kandidátovi dotaz, zda by bylo možné, např. přidáním vhodných programů, pokusit se zjistit i další frekvenční údaje o rodu, čísle a životnosti a zda by bylo možné při zařazení

sémantického analyzátoru provádět také analýzu závislosti deklinačních tvarů na slovesech z hlediska jejich sémantiky.

Jelínkovu práci jednoznačně doporučil k obhajobě.

Mgr. Tomáš Jelínek odpověděl na posudky oponentů. Na výhradu, že chybí vyrovnání se starší literaturou namítl, že v práci je zařazena jedna kapitola, v níž rozebírá relevantní data z jediné publikace na srovnatelné téma (Kvantitativní charakteristiky současné češtiny). Na námítku doc. Lopatkové, že chybí zasazení do kontextu současných bohemistických prací a vyrovnání s teoretickými přístupy k dané problematice, Tomáš Jelínek odpověděl, že omezení teoretického výkladu nebylo opomenutí, nýbrž rozhodnutí. Podrobný rozbor publikovaných prací o pádu a různých pojetí syntaxe v české a světové literatuře by jen odváděl pozornost od vlastní práce, stručný rozbor by nemohl zachytit nuance jednotlivých syntaktických systémů. Diskuse o pádu, kterou doc. Lopatková zmiňovala, se věnovala primárně tzv. „celostnímu“ významu pádů, což je pro tuto práci zcela irelevantní.

Na otázku ohledně využití existujících jazykových zdrojů (např. slovníku BRIEF) odpověděl kandidát, že využíval primárně seznamy vlastností slov z projektu automatické morfologické disambiguace založené na lingvistických pravidlech, což i uvádí v disertační práci. Konkrétně slovník BRIEF nevyužíval z důvodu spolehlivosti.

K další otázce kandidát poznamenal, že pořadí opravných pravidel je v práci uvedeno, možná ale ne dostatečně přehledně. Nejprve se aplikují pravidla pro opravu závislostní struktury souvětí (věty a spojky), poté se procházejí všechna slova ve větě od prvního po poslední a pravidla se spouštějí, když je nalezena chybná struktura.

Poslední otázka doc. Lopatkové se týkala úspěšnosti pravidel při aplikaci na optimálně nastavený parser. Tomáš Jelínek uvedl, že čísla, která uvádí doc. Lopatková jako nejlepší výsledek MST parseru, se týkají „unlabeled accuracy“, takže rozdíl použitého a optimálního nastavení zdaleka není tak veliký. V tabulkách ale představil i úspěšnost měřenou na parseru s výrazně lepším nastavením, kde se pravidla projevila méně, ale celková úspěšnost byla uspokojivá.

Na dotaz prof. Oldřicha Uličného ohledně prostého akuzativu ve funkci atributu uvedl Tomáš Jelínek dva příklady z korpusu a sdělil, že jde o zcela marginální jev. Většina případů započítaných do této kategorie jsou adordinační konstrukce typu „pan Novák“.

Ohledně nevhodného užití termínu *forma* připustil kandidát, že mohlo dojít k určitému zmatení pojmů, které bylo dáno využitím termínu *forma* v kontextu počítačové lingvistiky (kde se běžně užívá v jiném významu než v případě obecnělingvistickém).

Ohledně pojmu *pád* T. Jelínek namítl, že nepočítal *deklinační tvary*, ale abstraktní hodnoty gramatické kategorie tak, jak je interpretovala automatická morfologická disambiguace. Kandidát uznal, že pojem mohl být definován explicitněji, i když jeho definice je v práci uvedena.

Na další otázku kandidát sdělil, že technicky je poměrně triviální doplnit frekvenční údaje o rodu, čísle a životnosti, neučinil tak především z důvodu přehlednosti a srozumitelnosti tabulek. Možnost zařazení sémantického analyzátoru je problematičtější, protože neexistuje kvalitní analyzátor pokrývající dostatečné množství sloves. Pokud by takový analyzátor byl k dispozici, je možné ho do anotace zapojit.

Oba oponenti prohlásili, že odpovědi kandidáta na jejich výhrady uspokojivým způsobem reagují na námítky a že další už nemají.

Poté zahájil prof. František Čermák diskusi.

Diskuse:

Prof. Panevová poznamenala k námítkám doc. Lopatkové, že byly zčásti dány momentem zklamaného očekávání, kdy název práce odkazoval k teoretickým otázkám, zatímco obsah práce patří spíše ke kvantitativní lingvistice. Kandidát měl už do názvu zařadit pojem „kvantitativní“ nebo „frekvenční“ a explicitněji zdůvodnit, proč teoretickou část omezil.

Doc. Petkevič položil doc. Lopatkové (řečnickou) otázku, jak rozsáhlá má být „správná“ disertační práce a jaký podíl má mít vyrovnání s publikovanou literaturou ve srovnání s „vlastní“, inovativní prací. Uvedl příklad O. N. s disertační prací o 511 stranách, která podle něj kladla neúměrné nároky na čtenáře a byla právě kvůli rozsáhlému komentování publikované literatury příliš roztříštěná.

Prof. Kučera poznamenal, že by se kandidát vyhnul většině námitek, kdyby byl v práci explicitnější a přesněji zdůvodnil, co a proč uvádí.

Dále se prof. Panevová kandidáta otázala, proč počítá dohromady jednotlivé předložkové pády bez rozlišení předložky. Kandidát odpověděl, že je to nutné pro přehlednost tabulek, ale

jevy jsou podrobně kvantifikovány v předcházející části, takže všechny dílčí frekvence syntaktických funkcí s jednotlivými předložkami lze v práci dohledat.

Prof. Čermák poznamenal, že by bylo vhodné uvést i rozdělení podle mluvnického čísla. Po krátké diskusi se účastníci shodli, že statistické údaje s větším množstvím parametrů by bylo možné uvést v případné publikaci v přílohách.

Doc. Petkevič se otázel doc. Lopatkové na nejlepší parser používaný v současné době na ÚFAL MFF UK. doc. Lopatková uvedla, že se používá MALT parser v nastavení, které dosahuje úspěšnosti cca 86 % (ale neví, zda „unlabeled“ nebo „labeled“ accuracy). Doc. Petkevič se má v tomto směru obrátit na Daniela Zemana z ÚFAL MFF UK.

Vyhlášení výsledku tajného hlasování: Komise navrhla udělit titul doktor (Ph.D.).

Zapsala: Milena Hnátková

Podpis předsedy komise: