Posudek habilitačního spisu PhDr. Kláry Osolsobě, Dr.

Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny

Rozsáhlá práce dr. Osolsobě se zaměřuje na meze a možnosti (už takřka okřídlené spojení v jazykovědných kruzích!) automatického tvoření slov sufixací včetně morfologických alternací. Práce, opírajíc se o teoretický popis české morfologie, se věnuje jedné z jejích nejsložitějších oblastí, a to pro potřeby a z hlediska automatického počítačového zpracování češtiny propracovanými počítačovými nástroji a rozsáhlých datových základen – jazykových korpusů. K opodstatněnosti teoretických východisek autorky se nebudu vyjadřovat, neboť nejsem odborníkem na morfologii češtiny, to přenechám skutečným odborníkům na tuto problematiku. Zaměřím se stručně spíše na počítačové zpracování morfologie jako obzvláště rozsáhlé a spletité oblasti češtiny ve světle možností nabízených současnými počítačovými nástroji v rámci počítačového zpracování přirozeného jazyka, konkrétně na autorčinu schopnost exaktní algoritmizace derivace a autorčinu roli v brněnském přístupu k automaticky zpracovávané morfologii češtiny.

Jak známo, autorka spolupracovala na algoritmickém popisu české morfologie a strojovém morfologickém slovníku češtiny. Její pojetí je základem morfologického analyzátoru *ajka*, který se používá především – na úrovni morfologické analýzy – při automatické morfologické analýze rozsáhlých brněnských korpusů češtiny. Analyzátor *ajka* je známým konkurentem pražské morfologické analýzy autorsky stále spojované s jejím původním autorem Janem Hajičem.

V posuzované práci autorka předložila velké množství exaktně popsaných a podrobně komentovaných derivačních pravidel pro tvoření deverbativ: substantiv a adjektiv. Oproti morfologům působícím v předpočítačové době má dr. Osolsobě obrovskou výhodu v tom, co dnes činí matematickou/počítačovou lingvistiku odvětvím v rámci lingvistiky zcela nenahraditelným, neboť právě exaktními metodami matematické lingvistiky je možné explicitně ověřovat správnost teoretických východisek a závěrů a v podstatě jakýchkoli tvrzení o jazyce, jež se opírají o formu, tedy o něco materiálního, čeho se mohou zmocnit matematické a počítačové metody. A právě pojetí derivační morfologie v kontextu možností poskytovaných matematickými a informatickými metodami na jedné straně a korpusovými daty na straně druhé je silnou stránkou dr. Osolsobě – tím, čím Klára Osolsobě po mém soudu výrazně přispěla k rozšíření našich znalostí o zkoumaném tématu. Její práce je pěknou ukázkou toho, jak by si dnes měl počínat (počítačový) lingvista zpracovávající nižší roviny jazykového popisu (tedy nikoli například sémantiku či pragmatiku, ač pochopitelně i tam je maximální exaktnost neobyčejně žádoucí): Osolsobě formuluje exaktní substituční pravidla derivace, podrobně popisuje různé morfologické alternace a testuje svá pravidla na rozsáhlých korpusových datech. Právě verifikaci dnes umožňuje exaktně pojatá lingvistická bohemistika, tedy něco, o čem mohli přední lingvisté předpočítačové doby jen snít.

Nebudu rozebírat jednotlivá substituční pravidla obsažená v práci, uvedu pouze obecné závěry učiněné po podrobném pročtení práce. Na dr. Kláře Osolsobě tedy oceňuji tyto vlastnosti projevené v předložené práci:

- schopnost formulovat derivační substituční pravidla naprosto exaktně (to vůbec není něco samozřejmého: tato schopnost totiž řadě lingvistů zoufale schází!) tak, že je možné je hned implementovat nástroji automatického zpracování přirozeného jazyka; tedy – obecně řečeno – schopnost exaktně přemýšlet o české morfologii
- schopnost náležitě a efektivně analyzovat korpusová data (právě efektivně zkoumat rozsáhlá data je obtížné, neboť je například obtížné klást korpusu vhodné a zkoumanému problému přiměřené dotazy!)
- schopnost zobecňovat a formulovat právě tolik pravidel, kolik je zapotřebí (jakkoli celkové množství pravidel působí rozdrobeným dojmem)
- schopnost exaktně ověřovat pravidla na reálných korpusových datech, což dnes pro lingvisty představuje nesmírnou výhodu
- schopnost správně nahlédnout dva velké problémy automatického zpracování jazykových dat: *přegenerování* a *podgenerování*, přičemž autorka se snaží minimalizovat obě neřesti
- schopnost předvést ve světle velkých korpusových dat, jak takřka netušeně je čeština ve zkoumané oblasti složitá (ukažte mi podobný jazyk!).

Autorka si výtečně uvědomuje možnosti skýtané lingvistickým softwarem a v tomto smyslu lze říci, že je velmi dobrou matematickou lingvistikou, neboť spojuje hluboký a poučený vhled do jazyka, zde češtiny, se schopnostmi exaktního myšlení o jazyce konkrétně vyjádřeného v přesné algoritmizaci a testování vlastních algoritmů. Opakuji: jemná lingvistická analýza spolu s matematicky přesným myšlením není vlastní celé řadě lingvistů.

Obecně lze říci, že právě promyšlené zpracování morfologie je na brněnském pracovišti na velmi dobré úrovni, je to něco, na čem se dá v budoucnu stavět: dobré lingvistické pojetí, jež však chápe potřeby exaktní algoritmizace, se tu snoubí s další silnou stránkou brněnské počítačové lingvistiky: s vynikající úrovní programování. Oboje tak vede ke kvalitním jazykovým datům (morfologický slovník a další slovníky, databáze a korpusy) a k efektivním počítačově-lingvistickým softwarovým produktům. A autorka k vysoké úrovni matematické lingvistiky v Brně přispívá zásadní měrou.

Velmi též oceňuji, že autorka si je nadmíru dobře vědoma problému homonymie (v daném případě homografie) prostupujícího všechny plány přirozeného jazyka. Až počítačová éra nám totiž umožňuje nahlédnout tento problém v celé jeho šíři: vidět tuto šíři nám alespoň v oblasti morfologie umožňuje právě morfologický analyzátor opírající se o morfologický slovník.

Práce obsahuje mimo velké množství bohatě komentovaných pravidel i mnoho tabulek a také – zejména v závěru – grafy sufixálního tvoření deverbativ. Rozhodně se nečte snadno, ale látka je svou povahou tak složitá, že se tolika pravidlům nelze vyhnout. Publikovaná pravidla jsou navíc ověřena na velkých korpusových datech.

Mimo pravidla bych ještě rád vyzdvihl softwarový nástroj *Deriv* (kapitola VI.), na jehož vývoji se autorka zásadní měrou podílela. Velmi oceňuji, že tento nástroj je spojen s internetovým prohlížečem *DebDict* i s korpusem (dosud SYN2000). Právě organické

propojení různých datových zdrojů i softwarových nástrojů násobí možnosti efektivní práce s jazykem a lze je jen vítat.

Součástí práce je též popis derivačního slovníku deverbativ analyzovaných typů s příslušným webovým odkazem (kapitola IX).

Mám jeden konkrétní dotaz: v poznámce č. 118 na straně 58 autorka vychází z předpokladu, že „…tvary pasivního participia tvoří pouze slovesa přechodná, …". To však po mém soudu není pravda: deverbativa *narozen, dotázán, ublíženo* nejsou utvořena od sloves přechodných. Chci se tedy zeptat, jak to autorka s tou tranzitivitou myslí.

Obecně konstatuji, že autorce se podařilo splnit cíle vytčené mj. na začátku kapitoly IV: vytvořit formální popis derivace deverbativ a otestovat tak možnosti a meze formalizace (vždyť jazyk je nepravidelný, obtížně beze zbytku algoritmizovatelný!) a vskutku doplnit dosavadní popisy o dosud nezaznamenané případy. Mám za to, že velmi **exaktně pojatá práce**, která si však nelibuje v pouhých pozitivisticky pojatých soupisech, nýbrž vykazuje hluboký lingvistický vhled autorky do české derivační morfologie, **je dobrým základem pro úspěšné habilitační řízení autorky. Její habilitační práce splňuje požadavky standardně kladené na úroveň habilitačních prací v daném oboru**. Autorka má výtečné předpoklady velmi kvalitně pracovat v oblasti lingvistické bohemistiky, a to moderním způsobem, neboť dokáže využívat moderních metod umožněných rozvojem počítačového zpracování přirozeného jazyka. Takových odborníků, kteří se výtečně vyznají jak v lingvistice samé, tak v oblasti počítačového zpracování jazyka, není totiž bohužel mnoho: u drtivé většiny výrazně převládá jedna z uvedených oblastí.

V. Petkevič

V Praze dne 3. 3. 2012

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