Oponentský posudek práce

Formalizace systému české morfologie s ohledem na automatické zpracování českých textů

předložené k disertačnímu řízení na

Filosofické fakultě University Karlovy

RNDr. Jaroslavou Hlaváčovou

Obecné poznámky k zadání disertační práce a k obtížnosti zadaného úkolu

Základním úkolem předložené disertační práce bylo navrhnout a (pokud možno) i prakticky využít systém vzorů pro skloňování a časování českých slov vhodný pro zpracování českých ohebných slov na počítači (jak analýzu, tak syntézu slovních tvarů). Navržený systém by měl přitom být alespoň v některých parametrech lepší než všechny systémy již existující.

Složitost českého tvarosloví je přitom všeobecně známá a je jednou z překážek (byť asi ne překážkou největší) opravdu spolehlivého automatického zpracování českých textů. Mezi nejdůležitější úkoly v oblasti přípravy zdrojů pro základní i aplikovaný výzkum (psaného) jazyka patří přitom vytváření morfologické anotace rozsáhlých jazykových korpusů (např. korpusů řady SYN2000 Českého národního korpusu). Anotací se zde rozumí postup (a jeho výsledek), při kterém je každému slovnímu tvaru přiřazen základní slovní („slovníkový“) tvar (tzv. lemma) a značka určující morfologické kategorie (např. pád, rod, číslo, stupeň, čas, …) tohoto slovního tvaru. Anotace je přitom velmi podstatná podmínka pro jakékoliv praktické využití korpusu, ať už se jedná o vytváření slovníků, o lingvistický výzkum (který zásadním způsobem závisí na možnosti v korpusu vyhledávat podle abstraktních kritérií, tedy na úrovni anotace, nikoliv jednotlivých slov), či o učení a/nebo testování jazykových aplikací informačních technologií. Pro velké korpusy je ovšem prakticky neproveditelné anotování "ruční". Autorka se proto velmi důkladně věnuje právě otázkám morfologické analýzy korpusů, srovnává existující přístupy a navrhuje nové.

Vlastní posudek

Předložená disertační práce se z neformálního pohledu skládá ze dvou základních částí. V první z nich, kterou zhruba pokrývají kapitoly 1 – 6, je představena řada teoretických úvah a poznatků o českém tvarosloví a o tom, jak by měl vypadat jeho popis pro potřeby počítačového zpracování českých slovních tvarů v nejobecnějším kontextu i v konkrétním kontextu tvarosloví užitého pro potřeby anotace jazykových korpusů (zde se jedná zejména o repertoár jednotlivých morfologických kategorií a o jejich zachycení ve značce, která je každému slovu v korpusu přiřazena). Po „přechodové“ kapitole 7 zabývající se morfologickým slovníkem je pak v druhé části, pokryté kapitolami 8 – 13, navržen systém konkrétních vzorů (a okrajově i jiných metod) sloužících k zachycení různých flektivních typů českých slov. Hlavní linií práce je tak návrh systému, který je schopen popsat tvary českých slov, a to jak na úrovni teoretických úvah, tak na úrovni konkrétního rozpracování.

Při detailnějším pohledu je práce rozvržena do třinácti kapitol (první z nich je úvodní), závěru a doprovodných textů (seznamu literatury, přehledu morfologických kategorií a jejich možných hodnot a rejstříku).

První, úvodní kapitola představuje základní pojmy z oblasti automatického zpracování morfologie češtiny a ze značkování českých textů. Jako základní teze celého přístupu je zde uvedeno *„Zlaté pravidlo morfologie"*, které je definováno jako požadavek, aby byl každý slovní tvar jednoznačně popsán dvojicí ***<lemma, značka>***. Disertantka zde motivuje své rozhodnutí *Zlaté pravidlo morfologie* splnit tím, že popis slovního tvaru rozšiřuje na trojici ***<lemma, značka, mutace>***, kde „mutací" se rozumí další (doplňkový) popis vlastností slovního tvaru o údaje, které nejsou obsaženy ve značce (poněkud nešťastné je ovšem to, že tímto postupem se – přísně formálně vzato – *Zlaté pravidlo morfologie* plnit nezdaří …). Ocenil bych, kdyby při ústní obhajobě disertantka uvedla zdroj Zlatého pravidla nebo zdůvodnění, proč je formulovala, a uvedla motivy, proč se jím ve své práci řídila (žádná taková informace bohužel v práci není).

Druhá kapitola se věnuje problematice lemmatizace, tj. přiřazování „slovníkové" podoby ke konkrétnímu tvaru. Jedná se o téma jednoduché jen na první pohled, autorka uvádí řadu motivačních úvah v případech, kdy rozhodnutí není přímočaře jasné (např. u pravopisných variant typu *diskuse/diskuze*, u reflexiv tantum, např. *smát/smát se*, u záporných tvarů adjektiv – typ *nehezký*, u částí frazémů – lemma slova *krážem* aj.), i teoretický základ rozhodnutí, která v konkrétních případech přijala. I když by se na první pohled mohlo zdát, že jde o nemotivovaný "přílepek" k práci, která se zabývá jiným tématem, je realita právě opačná: jak se (nikoliv překvapivě) ukazuje, pro úspěšný popis morfologie je otázka volby lemmatu zásadní podmínkou.

Třetí kapitola popisuje dopodrobna problematiku výše uvedených mutací a teoretická rozhodnutí, která autorka v konkrétních případech přijala.

Rozsáhlá čtvrtá kapitola se věnuje problematice morfologických kategorií a jejich hodnot, tj. teoretickému pohledu na vymezení centrální problematiky práce: jaké morfologické kategorie lze (či: má smysl) pro jednotlivé slovní druhy (které jsou ovšem samy kategorií) určovat a jakých mohou tyto kategorie nabývat hodnot. Tato kapitola je napsána velmi důkladně (snad až na odkaz do tabulky 4.1 na třetím řádku zdola na str. 27, který je matoucí – má zřejmě jít o odkaz na tabulku 4.2, a na několik dalších, již ale jen drobných formulačních či jiných nepřesností) a podává vyčerpávající a bez nadsázky skvělý přehled o problematice české morfologie. Zejména v oblasti mutací (ale i jinde) navíc přináší řadu velmi dobrých detailních řešení pro „klasické" i nové problémy.

Pátá kapitola (velmi stručná, o rozsahu pouhých 2 stran) se věnuje speciální problematice podmiňovacího způsobu; osobně bych dal přednost tomu, aby její obsah byl přičleněn ke kapitole předcházející, ale její samostatné vyčlenění nechápu jako chybu.

Obsahově obzvláště závažná je kapitola šestá, která se věnuje problematice slov, jež disertantka označuje jako „složeniny", byť podle mého názoru jde o termín v českém názvosloví mírně zavádějící a dal bych přednost tomu, aby byl zvolen termín jiný (např. „agregáty"): nejedná se totiž o slova složená (kompozita, např. *tmavomodrý*, *světoobčan*), ale o „slepence" slov z různých slovních tříd, které si zachovávají – právě na rozdíl od slov složených – vlastnosti obou svých částí: jde o slova jako *naň, zač, očpak, viděls, včeras*. Kapitola přináší výborné myšlenky a ještě lepší řešení této velmi složité problematiky, vytknout v ní lze jen věcně a logicky nevalidní tvrzení, že z faktu, že v příkladu (83) na str. 61 je první složka složeniny v nominativu, vyplývá, že implicitní *jsi* zde nevystupuje ve funkci pomocného slovesa (že toto implicitní *jsi* zde neplní funkci pomocného slovesa, je sice pravda, ale jde o samostatný fakt, který nevyplývá z ničeho, a zejména ne z uvedeného

nominativu, o čemž svědčí možný protipříklad *Věrnýs nikdy nebyl,* kde je první složka slova *věrnýs* rovněž v nominativu, avšak implicitní *jsi* slovesem pomocným je), a zřejmý omyl v tabulce 6.3 na str. 64, kde je zájmenu *to* (v tomto tvaru) přiřazena hodnota čísla *S/P* (tedy možnost, že jde o singulár nebo plurál) místo zřejmého nepochybného *S* (singuláru). Tato kritika však nijak neumenšuje kvality této kapitoly, ve které disertantka dokázala schopnost samostatně vědecky zpracovat velmi složitou problematiku.

Sedmá kapitola je, jak již bylo řečeno, věnována morfologickému slovníku, tedy v podstatě seznamu všech tvarů českých slov a jim připsaných morfologických kategorií – řeší se zde v nejširším smyslu otázky zachycení takového seznamu a jeho správy (která vzhledem k rozsahu není nijak triviální záležitostí).

Kapitoly 8 – 13 popisují konkrétní vypracování vzorů a značek pro jednotlivé slovní třídy. V posudku se jimi nebudu zabývat jednotlivě, z obecného stanoviska jde o souhrn kapitol, kde disertantka prokázala schopnost samostatně vědecky pracovat: na základě empirického studia českých slovních tvarů, jejich podrobné analýzy a z ní vyplývající kritiky předchozích přístupů je zde vypracován systém flektivních vzorů pro všechny ohebné slovní tvary (s tím, že stupňování příslovcí je zahrnuto do kapitoly o přídavných jménech) a jsou přesvědčivě předvedeny jeho přednosti před přístupy dosavadními. Jediným drobným nedostatkem je snad absence alespoň krátkého historického úvodu (první systémy procedur morfologické analýzy češtiny vznikly daleko dříve než analýzy J. Hajiče a R. Sedláčka, které disertantka cituje, např. velmi rozsáhlý morfologický analyzátor češtiny byl vypracován na MFF UK na přelomu 70. a 80. let minulého století J. Dlouhým). Kromě „čisté" morfologie je v několika případech popsáno i odvozování (derivace), což je přínosem rozsahu množiny popisovaných slovních tvarů.

Závěr je stručná, ale výstižná přehledová kapitola popisující přínosy práce.

Z formálního hlediska je předložená práce vypracována velmi pečlivě, objevil jsem jen velmi malé množství překlepů (na str. 19 *vlasťovka* místo *vlašťovka*, na str. 24 je zřejmě „překlep v překlepu" – řetězec *ořejkeoz* má o jedno písmeno více, než by mít měl, na str. 112 *adjektnivními* místo *adjektivními*, na str. 129 *můsíme* místo *musíme*).

Závěrem mám tedy to potěšení konstatovat, že dle mého názoru předložená disertační práce požadavky kladené na práci toho druhu zcela splňuje, svým praktickým významem (nová morfologická analýza češtiny) je pravděpodobně dokonce převyšuje. Jejím autorstvím disertantka plně prokázala schopnost samostatné vědecké práce na vysoké úrovni, a proto doporučuji, aby Universita Karlova udělila RNDr. Jaroslavě Hlaváčové titul

Philosophiae Doctoressa

ve zkratce **Ph.D.**

V Praze 1. máje 2009

karel oliva