~~Oponentský~~ posudek diplomové práce

Morf.em.at.ic.ká strukt.ur.a sou.čas.n.é češ.t.in.y
Od lingvistické teorie k automatickému počítačovému zpracování

Jiří Lebeda

Předložená diplomová práce se zabývá morfematickým rozborem češtiny z velmi širokého spektra hledisek. Autor člení svou práci do deseti kapitol, z nichž nejvýznamnější jsou kapitoly *6. Metoda metodika komplexní morf(emat)ické analýzy*, *7. Morfematická gramatika a morfematicky orientovaná slovotvorba* a *8. Návrh efektivní morfematické syntézy*, ale velmi zajímavé jsou i ostatní kapitoly 0–5 a 9. Každá kapitola je na svém konci opatřena výstižným shrnutím. Práce mimoto obsahuje *Résumé, Seznam klíčových slov, Literaturu, Seznam obrázků* a tabulek a pozoruhodnou *přílohu*.

Celá práce dokládá vynikající nejen lingvistické vzdělání autora a jeho obrovský přehled o historické i současné lingvistické bohemistice včetně nejmodernějších možností dnešního počítačového zpracování češtiny. Autor je velmi podrobně obeznámen se starší i nejnovější literaturou, jež bezprostředně i vzdáleně souvisí s jeho ambiciózním tématem, kterým je zpracování morfematiky češtiny na základě slavné práce Eleonory Slavíčkové *Retrográdní morfematický slovník češtiny s připojenými inventárními slovníky českých morfémů kořenových, prefixálních a sufixálních* vydané v roce 1975 (dále RMSČ). Autor přinejmenším vyhověl *Zásadám pro vypracování diplomové práce* uvedené v *Zadání diplomové práce* a skutečně vypracoval komplexní metodiku morfematické segmentace češtiny, která vyhovuje nárokům lingvistické teorie i potřebám počítačového zpracování češtiny. Autor vychází z uvedeného dosud nepřekonaného morfematického slovníku a konfrontuje jej s největšími reprezentativními i nereprezentativními korpusy současné češtiny řady SYN. Výsledkem je vynikající, široce pojatá analýza velice komplikované morfematiky češtiny.

Začnu hlavními kapitolami. V kapitole 6. nazvané *Metoda metodika komplexní morf(emat)ické analýzy* se autor kriticky zabývá ruční, počítačovou a kombinovanou metodou morfematické analýzy a řadou souvisejících prací, kde nešetří kritikou a zároveň projevuje vynikající obeznámenost s tématem. S velkou znalostí věci zvažuje výhody a nevýhody ručního a počítačového přístupu a zatím stručně navrhuje, jak při zpracování morfematické analýzy postupovat v budoucnu: chce vyjít z RMSČ, uvést jeho nedostatky na pravou míru, a navrhuje (to ale až v kap. 8) efektivní (vlastně algebraický) systém zpracování morfematiky plně implementovatelný na počítači.

V kapitole 7. nazvané *Morfematická gramatika a morfematicky orientovaná slovotvorba* autor předvádí složitost morfematiky češtiny a srovnává přitom své hlavní zdroje: počítačově zpracovaný RMSČ a velké korpusy současné češtiny řady SYN z velkého množství hledisek. Údaje poskytnuté v desítkách tabulek obsahujících data z RMSČ a z korpusů řady SYN jsou – podrobněji jsem stačil prostudovat aspoň jejich část – nesmírně zajímavé a poskytují velmi dobrou představu o složitosti českého morfematického i morfologického (včetně slovních druhů) systému. Musela to být obrovská práce, klobouk dolů! Za všechny z obrovského množství zajímavých relevantních údajů bych rád uvedl údaj v pozn. 85 na s. 88 o počtu substantiv, jejichž tvary v korpusovém úzu realizují celé paradigma: je jich pouze 413! Jistě jsou to v drtivé většině ta nejfrekventovanější. V souvislosti s tabulkami bych měl pár dotazů:

- Jak autor zjišťoval propria v SYNČNK, když nejsou značkována jako taková?
- Jakou slabikovou segmentaci autor použil?
- Co to jsou homonyma (v odst. 7.2.1.3 na s. 90)? Lexikální typu *topit*, tvarová typu *zvířeně*, *tancích*, *divizně* apod., systémová v paradigmateh např. při pádové synkrezi? Jak je tomu v této souvislosti i u jiných údajů s rozlišením *typů* a *tokenů* (uvedená čísla se aspoň v SYNČNK týkají evidentně tokenů)?
- V odst. 7.2.1.6 je uveden poměrně vysoký počet 2339 lemmat RMSČ neobsažených v korpusu SYN2009PUB, 758 pak v korpusu SYNČNK (1,2 mld korpusových pozic). Je jich dost a jsou asi zastaralá. Jaká to jsou, prosím?

V kapitole 8 předkládá autor efektivní morfematickou syntézu v podobě – jak říká – dynamického systému, řídě se přitom proslulou Ockhamovou břitvou. Autor nejprve kritizuje dosavadní přístupy: koncepci lemmatu v pražském systému (souhlasím, koncepce už dlouho volá po zásadní revizi!), tápání kolem hadačů neznámých řetězců v textu, pouze grafematický režim morfologických nástrojů. Autor kritizuje i délku vývoje pražské morfologické analýzy (stále zatížené prastarou a dnes už málo vyhovující koncepcí) i následné disambiguace, ač té se Lebedova práce týká jen okrajově (přesto by mne zajímalo, jak by ji autor dělal, aby byla výrazně lepší než dosud!). Rovněž tak kritizuje přístup brněnský. Oba systémy kritizuje mimo jiné proto, že se neopírají o morfematiku a autorovu koncepci, a možná že má pravdu. Celý svůj systém navrhuje a velmi výstižně a koncizně popisuje v odst. 8.2 na s. 132–136. Předkládá velmi efektivní a ekonomický algebraický způsob morfematického popisu slov včetně dědění a výjimek z něho. Zvlášť bych při obhajobě rád diskutoval o obsahu s. 136, zejména o prvním odstavci s návrhy, jak zefektivnit morfologickou a slovnědruhovou disambiguaci, která je mým *métier*. Celá

kapitola 8 je jedním z jader práce, kde se nejvýrazněji projevuje algebraická metoda zpracování morfematiky.

Měl bych se ještě stručně zmínit o kapitolách 0 až 5, které tvoří jakousi širokou preambuli k jádru práce v kapitolách dalších. I tyto kapitoly jsou obecně řečeno velice zajímavé, obsahujíce pozoruhodné, často náležitě kritické postřehy. Opět zdůrazňuji, že autor v nich projevuje širokou i hlubokou erudici, zná obrovské množství literatury a projevuje velké znalosti lingvistické bohemistiky, nejen morfematiky češtiny. V uvedených kapitolách se zabývá morfémem jakožto konstituentem mentální reprezentace mysli, mentálním slovníkem a jeho fungováním, volá po spřaženém zpracování psané a mluvené podoby jazyka. Rozebírá pojem morfému v dějinách světové a české lingvistiky a uvádí velké množství jeho pojetí. Už jen vyznat se v peripetiích, jimiž prošel pojem morfému od Komenského dodnes, a sepsat jeho vývoj je nesmírně obtížné, leč velmi záslužné! V kapitole 4 autor systematizuje poznatky z formální morfologie, morfematiky, morfonologie a morfotaktiky a uvádí obsáhlé repetitorium poznatků z těchto odvětví s nejrůznějšími definicemi a pojetími relevantních pojmů. V kapitole 5 se zabývá především lexikografickým zpracováním morfematiky, uvádí velmi užitečný přehled vybraných lexikografických přístupů k morfematice, a to v češtině, slovenštině, polštině, ruštině a ukrajinštině a němčině.

Musím konstatovat, že málokterá odborná práce mě v poslední době tak zaujala jako práce Lebedova. Velmi oceňuji **algebraický** přístup k řešení nesnadného problému morfologické analýzy a syntézy: z malého množství jednotek (morfému) skládat velké množství větších útvarů – slovních tvarů, popř. jejich kombinací. Je to, jak autor jasně dokládá, postup velmi efektivní. Tento algebraický přístup je vlastně hlavním poselstvím práce: vycházejte od morfémů a větší jednotky skládejte z nich, je to nesmírně ekonomické. Jistě, to bychom ale museli mít morfematiku pro češtinu zpracovánu. Lebedova práce zde ovšem učinila velký krok kupředu: nejen v tom, co je v předložené práci, ale i v tom, co je za ní, tedy v počítačovém zpracování RMSČ a korpusových dat, jak dokládá obsáhlá kapitola 7 plná tabulek.

Soupis související literatury je obsáhlý a dokazuje suverénní autorův přehled o zkoumané problematice.

Velmi doporučuji, aby autor vystoupil se svým pojetím na některém ze seminářů ÚTKL, ÚČNK nebo ÚFAL a pochopitelně na bohemistických konferencích.

Mám za to, že žádný budoucí výzkum morfematiky češtiny (a možná i automatické morfologické analýzy a syntézy češtiny) a žádná tvorba některých

užitečných automatických nástrojů při počítačovém zpracování češtiny, jako je například hadač, by neměly Lebedovu práci opominout.

Jazyk práce je velmi kultivovaný, po slohové stránce bezvadný. Autor projevuje své široké (nejen lingvistické) vzdělání též volbou – mimo suverénně zvládnutou terminologii – někdy i dost výjimečných slov, třeba těch řeckého původu, jako je *katexochén*, *proteovský* apod.

Skromný autor na celou svou velkou práci nečerpal žádné prostředky od státu ani nadace v podobě grantu, podobně jako asi Josef Jungmann před ním. Ukazuje tak, že nevzniká-li nějaké i větší lingvistické dílo, netkví problém v penězích, ale někde jinde.

Nedostatky

Po uvedených kladech konstatuji, že jedinou vadou na kráse celé práce je nemalý počet překlepů; ad usum autora jsem většinu z nich vyznačil do jednoho exempláře práce při její četbě. Z mé korektury je tedy možné vycházet, bude-li chtít autor svou práci **publikovat, což bych velice doporučoval**.
Dalším nedostatkem je skutečnost, že na to, abych mohl ještě hlouběji ocenit autorovu práci včetně podrobného studia zajímavých tabulek prezentovaných v kapitole 7, jsem neměl mnoho času, pouhých deset dní.

Autor výrazně překročil požadavky kladené na diplomovou práci. Jeho nesmírně vyzrálou práci nejen **doporučuji k obhajobě**, ale dokonce ji po jistých výše doporučovaných úpravách **doporučuji vydat knižně**. Rovněž **doporučuji, aby práce byla podkladem pro zahájení rigorózního řízení**.

V Praze 11. 9. 2010

V. Petkevič

doc. RNDr. Vladimír Petkevič, CSc.
Ústav teoretické a komputační lingvistiky
Filozofická fakulta Karlovy univerzity