Posudek disertační práce

RNDr. Jaroslava Hlaváčová

Formalizace systému české morfologie s ohledem na automatické zpracování českých textů

Školitel: doc. PhDr. Vladimír Petkevič, CSc.

Cílem posuzované disertační práce bylo vytvoření systému pro přesný popis slovních tvarů, který je prvním předpokladem úspěšného automatického zpracování jazykových dat. Tohoto cíle má být dosaženo důsledným uplatněním „zlatého pravidla morfologie", které říká, že každý slovní tvar by měl být v systému popsán jednoznačně. Systém přirozeného jazyka (češtiny) toto pravidlo narušuje existencí variant na úrovni slovních tvarů i celých paradigmat. Zavedení termínu mutace slouží k jednotnému popisu umožňujícímu vzájemně odlišit všechny případy porušující „zlaté pravidlo morfologie". Jednoznačný popis variant na úrovni lemmatu sleduje zavedení vícenásobných lemmat. Druhou stranou téže mince představuje problém „složenin" (některých typů spřežek). Navržené řešení využívá koncept vícenásobného lemmatu, který zabrání ztrátě relevantních informací a zjednoduší vyhledávání sourodých jednotek ve značkovaných korpusech. Druhá část práce je zaměřena na popis morfologického slovníku (systém vzorů pro tvoření slovních tvarů a některých pravidelným a produktivním derivací).

Práce je členěna do 14 kapitol. Po úvodní kapitole, v níž se uvádějí základní definice a východiska práce, následují kapitoly, ve kterých se řeší problémy spojené s lemmatizací (2), navrhuje se jednotný, jednoduchý a (což je s ohledem na aplikace pro korpusovou lingvistiky pozitivní) teoreticky nezatížený popis variant na nejrůznějších úrovních (3). Dále autorka podrobně rozebírá klasifikaci jednotlivých tradičních i netradičních morfologických kategorií a jejich hodnot (4). Kapitoly (5) a (6) jsou věnovány rozboru případů, kdy jedna se nerovná jedné, tedy některým typům grafického spřažení více slov různých slovních druhů (*„složenin"*).

Druhá část je věnována Morfologickému slovníku (7) a jeho organizaci (nově navrženým vzorům, kterým se podle jednotlivých slovních druhů věnují kapitoly 9-13).

V závěru autorka stručně sumuje: a) motivaci práce, již spatřuje v potřebě reagovat na zkušenosti s více než desetiletou praxí využití nástrojů automatické morfologické analýzy v nejrůznějších aplikacích, b) obsah práce a c) dosažené výsledky.

Vlastním přínosem práce je tedy:

a) navržení rámce pro systém přesného popisu slovních tvarů na základě nově definovaných kategorií, jejich hodnot a flektivních vzorů.

b) vytvoření nástrojů pro převod vzorů současně používaného „pražského systému" do nově navrženého systému vzorů.

c) prefixový guesser.

d) postfixový guesser.

e) návrh konkrétního schématu pro uchování morfologického slovníku v PML (Prague Markup Language) s ohledem na možnost zaznamenávat derivační vztahy.

Hodnocení:

- Autorka staví svoji práci na mnohaleté zkušenosti s využitím nástrojů automatické morfologické analýzy v praxi. Ačkoliv to v práci není explicitně řečeno, ale pouze naznačeno odkazy na „lingvisty neoblíbená technická řešení", hodnota aplikace v oblasti NLP není podmíněna výlučně teoreticky, ale též prakticky. V tomto rámci práce vychází z praxe a směřuje k jejímu vylepšení.

- Autorka prokázala, že je schopna zpracovat velmi různorodé a mnohdy protichůdné podněty.
- Autorka vytvořila konzistentní teoretický rámec, který podrobně popsala a navrhla i nástroje pro jeho aplikaci.

Připomínky: Slabinou textu předložené disertační práce je úroveň citací. Na prvním místě bychom vytkli nedostatky v seznamu použité literatury, v níž chybí a) odkazy k použitým korpusům, b) citace použitých nástrojů pro práci s korpusy, slovníky atd., c) některé práce k nimž se v textu odkazuje (např. nějaký text týkající se systému automatického morfologického zpracování slovenštiny – srv. s. 54 odst. 4), d) tištěné slovníky (přestože výkladové slovníky jsou zastaralé, stále jistou prestiž mají), e) gramatiky (autorka cituje pouze 2. díl akademické mluvnice a mluvnici Havránka a Jedličky). Největší nedostatek spatřujeme v tom, že necituje alespoň 1. díl akademické mluvnice, přestože se zabývá pravidelnými derivacemi tj. slovotvorbou, jíž je druhá část prvního dílu akademické mluvnice věnována. Autorka uvádí, že nebude v práci řešit různé lingvistické přístupy k řadě sporných a otevřených otázek. Přesto mohla dát najevo, že tento, jak se domníváme, legitimní postup je opřen o znalost syntetických prací, které ovšem nenabízejí odpovědi na otázky kladené praxí NLP, na něž disertační práce Jaroslavy Hlaváčové hledá odpovědi.

K jednotlivostem:

1) Nesouhlasíme s některými tvrzeními uvedenými v práci.

Na s. 8 se diskutuje lemmatizace zvratných sloves. Nesouhlasíme s tvrzením „*obě části zvratného slovesa od sebe mohou být* ***vzdáleny****, a to* ***libovolným*** *počtem slovních tvarů, dokonce na* ***obě*** *strany.*"

Na s. 16 se uvádí charakteristika variant s protetickým *v-* a s instrumentálovými spisovnými a nespisovnými koncovkami, přičemž autorka poznamenává, že varianta „*3 (vokýnky?) je podivná, neboť má nespisovný kmen a spisovnou koncovku.* Dle našeho názoru jde o jev, který není nikterak „podivný". Setkáváme se s ním na vyšší rovině např. syntaktické, kdy ke střídání kódů (spis./nespis.) dochází např. uvnitř jmenné skupiny („***ty*** *dobří kluci*").

Nesouhlasíme s protipříklady slov, která dle autorky „*nemohou protetické v- přijímat*" (pozn. 1 na s. 17). Lze vyvrátit doklady z internetu i mluvených korpusů.

Nesouhlasíme s tvrzením na s. 25, kde se říká, že jedině verbální substantiva na *-ní/-tí* mohou mít „*zvratnou částici*". Mohou ji mít i deverbativní adjektiva, která se mohou substantivizovat (např. ***Nehodící se*** *škrtněte*).

2) Upozorňujeme na místa, která by pro lepší orientaci čtenáře měla být upřesněna.

Jak máme rozumět poslední větě úvodního odstavce na s. 3 a k ní se vížící poznámce č. 4 ? „*Interpunkce je zpracována dostatečně, není proto důvod se jí dále zabývat.*" a „*Uvažujeme o vytvoření klasifikace funkcí jednotlivých interpunkčních znamének.*" Naráží autorka na problém, kdy je např. znak „..." (tečka) součásti slova (iniciálová zkratka) atd., nebo jde i o jiné případy ?

Na s. 30 v komentáři k obr. 4.1 se uvádí, že neurčité číslovky se netvoří pomocí „zneurčiťujících" přípon. Kam se řadí *koliksi*? Uvádí se též, že neexistuje **néodkud*, a opomíjí se, že existuje *od-ně-kud*.

3) Nabízíme k úvaze a doplňujeme.

K poznámce 1 na s. 21 upozorňujeme, že klasické morfologické práce rozlišují gramatické kategorie prvního (slovní druhy) a druhého řádu (slovnědruhově závislé kategorie). Slovní druh tedy jako gramatickou kategorii chápat lze.

Kladně hodnotíme autorkou navržené rozdělení kategorií určovaných při automatické morfologické analýze na globální a flektivní. Navrhujeme k úvaze, zda mezi kategorie globální nezařadit podobně jako vid sloves rod substantiv. V obou případech je splněn

požadavek definice globální mutace (vztahuje se na celé paradigma). Z lingvistického hlediska se rod substantiv někdy chápe jako klasifikační kategorie (srv. též str. 48 odst. 1).
Na s. 60 se v komentáři ke „slovesným složeninám" s *-s* za 2. os. sg. uvádí, že „*Přes vysokou produktivnost* (?produktivitu) *tvoření slovesných složenin však není jejich výskyt příliš častý.*" Počet výskytů je závislý na typu textu na konkrétním složení korpusu. Srovnání relativních frekvencí příslušného jevu ukazují, že v korpusu složeném z dialogických textů (Korpus soukromé korespondence - ksk) je frekvence příslušných složenin až 18x vyšší než v obecných korpusech.
Ačkoliv se v českých gramatikách uvádívá totéž, co v první větě 2. odst. na s. 124, totiž „*Podstatná jména slovesná se zakončením -í se odvozují z tvaru trpného rodu ...*", hovoří proti tomuto pojetí (resp. formulaci) fakt, že deverbativa na *-ní -tí* a) pasivní význam nemají a b) tvoří se i od sloves, která tvary příslušných participií (adjektiv jmenných) netvoří (např. *bytí, ležení, sezení, stání, ...*).

4) Upozorňujeme na některé formulační nedostatky a opomenutí.

Na s. 20 v posledním odstavci se odkazuje k tabulce 4.4 na s. 34, ve skutečnosti je tato tabulka uvedena až na s. 36.
Za přehlédnutí pokládáme, že se na s. 27 uvádí, že substantiva *kdo, co* mají „*substantivní skloňování*". Ve skutečnosti jde o skloňování zájmenné, zvláštní.
Na mnoha místech práce se autorka zmiňuje o řešeních, která jsou tak či onak lingvisticky „nekonvenční". Jedním z nich je i řazení pasivních participií mezi jmenné tvary adjektiv. Pokud jde o zdůvodněné řešení konkrétního popisu, je to v pořádku. Nicméně by bylo v takovém případě vhodnější vyhnout se formulacím jako: „*Slovní tvar* ***ukryt mohl být odvozen*** *jak z přídavného jména* ***ukrytý****, tak ze slovesa* ***ukrýt***." (Jde o užití min. času). V kapitole **4.2.1 Rod (GEN)** (s. 38) se explicitně uvádí, že rod je relevantní jenom pro „... *slovesa v příčesti činném*". To, že trpná participia jsou v systému zahrnuta mezi adjektiva, se ovšem v textu práce čtenář dozví až na s. 45.
Za přehlédnutí pokládáme na s. 12 komentář k příkladům (17) a (18): „*Zatímco v prvním příkladě jde o sloveso s* ***akuzativní*** *valencí, druhý příklad je intranzitivní.*"
Na s. 70 kap. **7.1.1.2. Postfixový guesser** nerozumíme formulaci: „*Čím delší zakončení, tím méně možností guesser nabídne,* ***ale tím více jich je zapotřebí.***"
Na s. 94 chybí ve výčtu na začátku stránky kód pro vzor `mtr-eu-e` (*patro*), který je uveden v tabulce na s. 95.
Na s. 109 se uvádí, že derivace typu `Do` (adverbia tvořená od adjektiv slovnědruhovou charakteristikou *-o*) se nestupňují. Řada adverbií na *-o* ovšem II. i III. stupeň tvořit může. Většinou jsou to sice adverbia, ke kterým (jak se uvádí na s. 114) existují varianty na *-e/-ě* nebo *-y* (*kolmo, kolmě, kolméji*). Předpokládá-li se, že v případě existence dublet (víc než jedno lemma) je pro stupňování relevantní tvar na *-e -ě* nebo *-y*, pak by to mělo být řečeno na příslušném místě explicitně. Je tomu tak opravdu ? (srv. *Je jim čím dál smutněji.*)
Na s. 119 – formulace vysvětlení rozdílu hodnot m a n v tabulce 12.4 je nepřesná.

Otázky k diskusi:

Na několika místech své práce autorka navrhuje řešit problémové případy „individuálně". Zmiňuje se např. o lexikalizacích, které vedou stručně řečeno ke změnám kategorií i hodnot v navrženém systému interpretací (lemmat, existence a konkrétních hodnot kategorií jako negace, stupeň atd.). Má autorka představu o tom, jak by se mělo postupovat, aby „jednotlivá řešení" byla konzistentní ? Konkrétně:
1. Ke kapitole **2.2.3 Slovní tvary „bez lemmat"** je třeba poznamenat, že zvolené příklady vyvolávají pochybnosti o navrhovaném řešení. Jedná se o návrh lemmatizace tvaru *bycha* („*pozdě bycha honit*"). V citovaném korpusu jsou dva výskyty tvaru instrumentálu (*bychem*) a několik tvarů „*bycha*" chybně označkovaných jako tvary akuzativní, ačkoliv se jedná o

s ak. homonymní tvary genitivní, které jsou výsledky transformace *„honit bycha > honění bycha"*. Ve slovnících (SSJČ, PSČ) uváděný nominativ (*bych*) v korpusu ovšem doložen není. Za sporné pokládáme návrh řešit tyto případy „individuálně". Jaká opatření se budou muset dodržet, aby přijatá řešení nebyla řešeními ad hoc ? Totéž platí i o návrhu řešit „jednotlivě" případy lexikalizovaného užití neboli dezaktualizace verbálních adjektiv z přechodníků přítomných (s. 26). Na s. 122 ve 2. odst. pod tabulkou se hovoří o pravidlu pro tvoření adjektivizovaných přítomných přechodníků. Bude přechod adjektivizovaných přechodníků (nestupňovatelných) v dezaktualizovaná adjektiva (? stupňovatelná) (srv. s. 26) řešen zdvojením lemmat na úrovni slovníku ? Má autorka nějaký (zřejmě otevřený) seznam lemmat (založený např. na korpusu)?

2. Na s. 28 kap. **4.1.2.5 Poddruh sloves** jsou jednotlivé typy definovány výčty. Jsou výčty úplné ? Jaká byla kritéria pro zařazení jednotlivých sloves ? Podle jakého klíče byla zařazena mezi modální slovesa *mívat, musívat, ...* a nebylo zařazeno např. *chtívat*?

3. Pro případy uváděné v kap. **4.2.7 Negace (NEG)** existuje termín **negativum tantum.** Autorka práce nechává rozhodování o tom, jak lemmatizovat dvojice, které se liší +- přítomností úvodního řetězce *ne-* na *„správci konkrétního slovníku"*. Druhá část posuzované disertace ke konkrétnímu slovníku odkazuje. Má autorka alespoň pracovní seznam negativ tantum ?

4. Na s. 42 v kap. **2.6 Stupeň (DEG)** se autorka poměrně stručně vypořádává s otázkou zařazení stupňování v mluvnici. Operuje s termínem „stupňovatelné lemma". Podle jakého klíče se budou na úrovni slovníku příslušná lemmata hodnotit?

5. Na s. 123 se v poslední větě uvádí, že od adjektiv z přechodníků na *-cí* se příslovce netvoří. Má autorka nějakou představu systematické lemmatizaci tvarů na *-cnější* (*nejalarmujícnější*) a propojení derivací na *-cně* (*otřásajícně*) a na *-cnost* (*strhujícnost*) s možnými fundujícími lemmaty ?

Ve druhé části práce věnované vzorům autorka podává přehledný návrh nové klasifikace vzorů. V řadě případů je navrženo obecné řešení, jehož realizace ale popsána není. Není patrné, podle jakého klíče se zachází s „jinými" na úrovni kategorie mutace a s výjimkami na úrovni navržených vzorů. Konkrétně:

1. Jaký je klíč pro zařazení jednotlivých případů mutací pod hodnotu „jiné"? Je zařazení mezi „jiné" opřeno o nějakou např. kvantitativní analýzu korpusových dat? V tabulce zachycující globální mutace (s. 36, tab. 4.4) se např. neuvádějí případy střídání *u-ou* (např. *hrouda-hruda, sluha-slouha, strouha-struha, ...*).

2. Na s. 83 se v 1. odst. hovoří o pravidelných derivacích posesiv na *-uv, -in*. Autorka píše, že derivace se uvádí explicitně u ženských vzorů. *„neboť u mužských životných vzoru lze přídavné jméno přivlastňovací utvořit zřejmě vždy."* Podle jakého klíče se řadí feminina k jednotlivým vzorům?

3. Tamtéž se v 5. odst. uvádí, že *„Všechny vzory podstatných jmen mají pravidelné flektivní mutace FMC=a se zakončením –ma.... – pány- pánama, stroji-strojema, ...* Explicitně se neuvádí u měkkých vzorů druhá (výrazně moravská) varianta *(hranicama, věncama, lyžama, mikulášama, stavebnicama, nožama, hercama, ...)*. Všechny příklady jsou ze SYN. Je zařazena ?

4. Na s. 86 kap. **9.2.2. STROJ** se neuvádí varianta sjx-2pl**P** pro substantivum *peníze* (? jiná). Patří do výjimek ? Podobně na s. 99 kap. **9.3.2. MUŽ** se neuvádí, že vzor mz1eFa *obyvatel* má flektivní mutaci v gen. pl. (srv. *1500* ***učitelů*** *x 1500* ***obyvatel***).

5. Na s. 97 autorka uvádí, že se nezabývá *„otázkou kodifikovanosti"* mutací koncovek nom. pl. vzoru pán. Platí totéž i o koncovkách dat. a lok. sg. všech vzorů životných maskulin ?

6. Na s. 98 se v tabulce 9.16 uvádějí *„Možná zakončení vzoru pán"*. Měla na zařazení k *„ostatním zakončením"* vliv analýza frekvence jednotlivých mutací v korpusech ? Pokud ne, podle jakého klíče autorka postupovala ?
7. Na s. 103 se uvádí vzor y23 pro skloňování 2. a 3. stupně adjektiv, u nichž dochází *„ke změně ve kmeni"*. Kolik adjektiv a která (seznam) se podle tohoto vzoru budou ohýbat?

Předložený text splňuje podmínky kladené na disertační práci. Doporučujeme jej tudíž k obhajobě.

PhDr. Klára OSOLSOBĚ, Dr.
FF MU BRNO