

## Oponentský posudek diplomové práce

Josef Toman: Automatická anotace angličtiny na tektogramatické rovině

### Obsah práce

Předložená práce je zaměřena na implementaci a vyhodnocení souboru automatických procedur zaměřených na zvýšení efektivity při značkování jednoho z korpusů vytvářených v Ústavu formální a aplikované lingvistiky. Práce je členěna následovně. Po úvodní kapitole následuje kapitola popisující Pražský závislostní anglický korpus. Třetí kapitola zmiňuje existující datové zdroje, které budou použity pro automatizaci značkování korpusu. Čtvrtá kapitola vysvětluje evaluační metriky zvolené pro zadanou úlohu. Pátá kapitola, která tvoří těžiště práce, uvádí jednotlivé moduly vytvořené pro předznačkování materiálu, včetně jejich detailního vyhodnocení. Šestá a sedmá kapitola se zabývá doplňkovými možnostmi, jak zefektivnit anotaci, konkrétně užitím automatických testů v anotačním prostředí a automatickým postprocessingem ručně anotovaných dat. Osmá kapitola předkládá vyhodnocení celé soustavy procedur jako celku. Závěrečná devátá kapitola shrnuje dosažené výsledky. Práce je psána česky, včetně seznamu literatury má 81 stran.

### Hodnocení

Nejprve je třeba uvést, že diplomantova práce skutečně urychlila a zefektivnila vývoj korpusu anglických tektogramatických stromů. Hlavní cíl byl tedy dosažen.

Práce je psána velmi srozumitelně a je přehledně strukturovaná. Na řadě míst je text doplněn ilustrativním materiálem, mj. příkladovými větami, vzorky externích dat a fragmenty tektogramatických stromů.

Činnost každého dílčího modulu pro automatické zpracování vznikajících lingvistických struktur je podrobně popsána. Je zřejmé, že autor při jejich implementaci pronikl do podstaty zpracovávaných struktur, podle potřeby hbitě přepíná v užívání tektogramatických a složkových stromů. Značná úspora anotačního úsilí byla umožněna také díky autorově snaze o integraci všech již existujících externích anotací, mj. struktur jmenných frází a textové koreference.

Kladně hodnotím autorův důraz na detailní vyhodnocení, které bylo z důvodu možného rozdílného tvaru stromů netriviálním úkolem a vyžadovalo nejprve vyřešení problému, jak jednotky srovnávaných stromů vůbec párovat. Vyhodnocení jednotlivých modulů je provedeno pomocí popsaných metrik ve formě tabulek, jejichž jednotné zpracování usnadňuje orientaci. Autor se snaží vždy analyzovat příčiny úspěšnosti nebo neúspěšnosti jednotlivých metod, přičemž některá z pozorování upozorňují na fundamentální otázky a problémy anotačního procesu. Mimo jiné jde o doloženou změnu chování anotátorů v závislosti na různém předzpracování dat (např. na str. 66).

Autor v textu zřetelně vymezuje vlastní roli v projektovém týmu. Zde je ale na místě připomenout, že kromě vývoje pravidel pro automatickou anotaci popsaných v předložené práci měl na starosti i všeobecnou technickou podporu anotátorů.

Po jazykové stránce je předložená práce na solidní úrovni. Na několika místech je ale popisnost textu poněkud nadbytečná, např. druhá polovina sekce 3.4. Nelze také přehlédnout občasná vybočení z odborného stylu („vcelku sympatický přístup“, „Nejprve jsem si řekl, ...“. „Číslo je to sice hezké, ale ... žalostně nízké“, „Stopy prozatím úplně odstavíme“, „Tomu odpovídá... mizerný výsledek“, „Systém se tvrdohlavě snaží nějakou spojku najít a výsledkem je nesmysl“, „Teprve potom se začalo pořádně pracovat ...“, „Absolutní čísla vypadají hezky, i když ztrácí své kouzlo...“, nadužívání slova „spousta“). Dále bych za nevhodný označil „anglický“ slovosled ve spojeních typu „BBH korpus“ či „PTB soubor“.

V textu lze nalézt malé množství gramatických chyb a překlepů („... skutečnosti mě vědly“, „Výsledky ... byly až malé odchylky totožné“) a terminologických nepřesností („neterminál s formou“, „apozice je reprezentována koordinací“).

Prezentaci číselných výsledků by podle mého názoru bylo možné vytknout, že hodnoty jednotlivých metrik jsou na některých místech předkládány se zavádějící přesností (např. úspěšnost pravidla s jednou chybou ze šesti pokusů zaokrouhlená na čtyři platné číslice).

Uvedené nedostatky ale hodnotím (s ohledem na celkový objem odvedené práce a na převažující prezentační kvalitu) jako drobné.

### Doplňující otázky

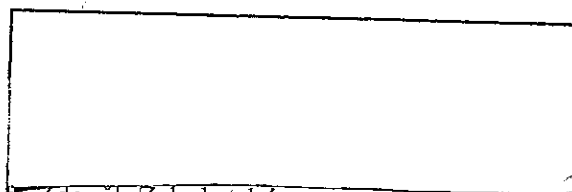
Rád bych diplomanta požádal o krátké vyjádření k následujícím otázkám.

- (1) Na straně 15 autor uvádí, že rozdělení na trénovací a testovací data zde ztrácí smysl a že cílem by mělo být co největší přetrénování modulů. Prosim o podrobnější vysvětlení této myšlenky. Vedle „starých“ a „nových“ dat použitých při evaluaci přece existují i data, která budou ručně anotována až v budoucnu a pro jejichž předzpracování je přetrénování nežádoucí.
- (2) V sekci 5.5 autor popisuje odstranění odkazů na stopy ve složkové struktuře, zatímco v sekci 5.12 popisuje naopak modul pro vytvoření nových uzlů pro některé typy nevyjádřených aktantů, které ovšem v zásadě odpovídají původním stopám. Nebylo by tedy vhodnější tyto stopy ve složkovém stromu a rekonstruované uzly v tektogramatickém stromu spíše propojit?
- (3) Preanotační pravidla byla konstruována na základě znalosti lingvistické podstaty dat. Bylo by možné automatizovat i hledání těchto pravidel?

### Závěr

Diplomant v předložené práci potvrdil, že dovede samostatně zpracovat náročné téma. Prokázal vhlad do podstaty problému i značnou programátorskou zdatnost. Práce splňuje zadání, prezentace dosažených výsledků je precizní.

V Rychnově nad Kněžnou, 30. března 2009



Zdeněk Zabokrtský  
Ústav formální a aplikované lingvistiky  
MFF UK, Praha