Oponentský posudek doktorské disertační práce

Václav Klimeš: Analytical and Tectogrammatical Analysis of a Natural Language

Doktorská disertační práce V. Klimeše si klade za cíl vytvořit automatický analyzátor pro analýzu na analytické a tektogramatické rovině, který pracuje s anotacemi použitými v PDT (proč je první odstavec v odd. 1.1 v uvozovkách?, navíc anglická formulace je velmi neobratná).
Práce se skládá ze tří částí, první je tvořena kapitolou úvodní, druhá část je věnována analýze na analytické rovině, ve třetí autor probírá analýzu tektogramatickou a čtvrtá kapitola shrnuje výsledky práce. V závěru najdeme rejstřík, seznam použité literatury a přílohu A. K práci je přiložen CD ROM obsahující text práce a příslušné přílohy.

Práce je psána anglicky, text je díky dlouhým větám a konstrukcím typu Czenglish nesnadno srozumitelný, viz např. na s. 28, 29, 31, 32, 50, 51. Autor má občas problémy s koreferencí, např. na s. 30 není jasné, k čemu odkazuje výraz *this parsing method*, podobně tak i na s. 31 a jinde. Korektura ze strany anglického mluvčího by velmi příznivě ovlivnila srozumitelnost práce.

Přínos práce:

a) byly vytvořeny nové nástroje pro práci s analytickou a tektogramatickou rovinou v rámci FGD, získané výsledky lze pokládat za originální.
b) vytvořené analyzátory jsou příspěvkem v oblasti parsingu, využívá se jich v ÚFAL pro anotování na obou zmíněných rovinách.
c) Výsledky lze pokládat za slibné pro další práci.

Problematické body a otázky:

a) V práci se nabízela možnost podívat se hlouběji na některé aspekty popisu jazyka, jenž pracuje s analytickou a tektogramatickou rovinou v rámci FGD. Autor tu měl možnost položit si otázku (s. 13-14), jak a nakolik je použitý teoretický rámec odůvodněn empiricky, tj. podložen korpusovými daty. Zajímavý by např. byl pokus provést validaci funktorů vůči korpusovým datům (ocenil bych, kdyby autor mohl na tuto otázku zareagovat při obhajobě).
b) Předkládaná práce se těmito teoretickými aspekty nezabývá, rámec FGD pokládá za pevně daný, z toho pak vyplývá její převážně technická povaha. Hlavní náplní je pak implementace již zmíněných nástrojů, což nepochybně omezuje její teoretický přínos.
c) Některé pojmy nejsou definovány nebo citovány, např. linguistic meaning" na s. 18, nebo „semantic part of speech" na s. 19., podobně i „inference of rules" na s. 31. Uvedené pojmy se ani neobjevují v rejstříku. Pojem „rule template" uvedený na s. 23 je charakterizován bez potřebného příkladu, na s. 24 sice najdeme příklad pravidla, které je instancí „rule template" ale vhodný ilustrativní příklad „rule template" by jistě přispěl ke srozumitelnosti výkladu.
d) Přehled parserů pro češtinu uvedený na s. 29-30 není v žádném případě úplný, autor by se měl lépe seznámit s existující literaturou. Kromě toho sám termín „parser" není v práci přesně definován a ani se neobjevuje v rejstříku (v něm je volně zmíněn „parsing" bez jakéhokoli odkazu na relevantní literaturu, i když řada prací jinde v disertaci citována je).

e) Na s. 31 se mluví o „unlimited number of tokens" – co se tím myslí, může takový případ vůbec nastat?
f) Na s. 43 se říká, že počet neprojektivních složek v češtině je nízký (low) (s odvoláním na Zemana 2005). Korektní by bylo citovat přesné počty. Kromě toho existují práce, které na základě korpusových dat říkají, že např. frekvence neprojektivních slovesných složek v češtině vychází na 20 %. Váhal bych bez dalšího říci, že tato frekvence je nízká.
g) na s. 64 a 65 se probírá příklad z obr. 3.3 – spojení na a-rovině *více než* je na t-rovině převedeno na *hodně*. V příkladové české větě jde o srovnávací obrat, který má evidentně jiný význam, než naznačuje zápis na t-rovině. Vzniká pak otázka spolehlivosti takových anotací, i když autor práce za chyby tohoto druhu patrně nenese přímou odpovědnost. Je-li jich ale více, získané výsledky mohou být problematické, zvlášť když se mají používat jako trénovací data. Nezdá se mi, že by si autor byl tohoto nebezpečí vědom a že by na ně jakkoli reagoval.

Hodnocení:
Hlavním výsledkem práce jsou 2 parsery a nástroj pro přiřazování syntaktických tagů (čemu?). V závěru na s. 91 autor mluví současně o parserech a metodách 1 a 2 a zjevně mezi nimi nerozlišuje, i když „parser" a „metoda" jistě nejsou synonyma. Autor uvádí, že přesnost získaných výsledků je pro oba parsery asi o 10 % nižší než u standardních parserů. Pokud jde o t-analýzu, zlepšení proti dosavadním nástrojům je, jak se v práci uvádí, 29 a 47 %. I zde autor mohl vedle sebe uvést hodnoty svých a předchozích výsledků, aby čtenář nemusel celková porovnání výsledků pracně dohledávat sám.
Pokud se nástrojů popsaných v práci používá pro usnadnění manuálního značkování při budování anotací pro uvedené 2 roviny, lze získané hodnoty jistě pokládat za příznivé. Otázkou, jak by tomu bylo u realisticky orientovaných aplikací v oblasti počítačového zpracování jazyka, se autor nezabývá.

Závěr:
Autor v práci prokázal, že dovede samostatně vědecky pracovat a řešit problémy v oblasti počítačového zpracování přirozeného jazyka. Vlastní provedení práce je poněkud nevyrovnané a jak jsme naznačili výše, trpí řadou nedostatků formálních (jazykových) i věcných (práce patrně vznikala ve spěchu?). Přes uvedené výhrady však předložená práce může sloužit jako podklad pro získání stupně Ph. D.

V Brně, 20. 8. 2006

Karel Pala
Katedra informačních technologií
Fakulta informatiky MU