Disertační práce Mgr. Jana Štěpánka „Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konsistence dat)“, MFF UK Praha 2006

Vyjádření školitele

Předložená disertační práce je souhrnem cenných poznatků o práci s velkými anotovanými soubory dat, o jejich ukládání, zpracování a kontrole. Týká se práce s anotovanými korpusy obecně a s Pražským závislostním korpusem (dále PZK) zvlášť. Nutnost získat praktické zkušenosti pro vývoj nástrojů pro práci s korpusy jak z oblasti informačních technologií, tak z teorie a empirie zpracování těchto dat vedla k tomu, že výsledky v podobě disertace mohly být předloženy až na konci 7. roku zprvu interního, poté kombinovaného doktorského studia. Mgr. Štěpánek si osvojil velmi důkladně, ale s kritickým nadhledem teoretický rámec pro anotování PZK (viz 1. kapitola disertace) – funkční generativní popis. V práci popsal a zhodnotil nástroje pro práci s anotovaným korpusem (TrEd, ntred, btred aj. v kap. 2). Popsal a vyhodnotil různé možné přístupy ke zpracování koordinačních konstrukcí jakožto jevu obtížnému pro všechny typy formalismů, včetně závislostního. Hledání efektivních synů a efektivních rodičů vyžaduje u těchto konstrukcí zvláštní ošetření a je mu v práci věnována velká pozornost. Mgr. Štěpánek navrhl procedury, jak prohledat i tyto konstrukce, pro něž dotazy sloužící k běžnému vyhledávání pomocí nástroje NETGRAPH by byly obecně v některých případech nemožné. Autor disertace se významně podílel na převedení PZK do nového anotačního schématu (PML), které zajišťuje propojení anotace na všech rovinách (morfologické, analytické a tektogramatické). Vytvořil a v praxi ověřil nástroje pro poanotační kontrolu dat (kap. 3). Kontrolované jevy roztřídil do typů, podrobně a s velkým pochopením popsal jednotlivé typy chyb a provedl evaluaci použití kontrolních mechanismů na datech PZK (verze PDT 2.0). Tam spočívá těžiště práce, nepochybně šíře využitelné pro jiné korpusy (v „annotation science“, jak se moderně začalo říkat).

Ve své práci doktorand prokázal hlubokou znalost lingvistické teorie a porozumění empirickým datům, stejně jako potřebnou erudici pro efektivní práci s korpusovými daty. Jím navržené a použité nástroje nesporně přispěly k významnému zkvalitnění PDT 2.0 a ukázaly cesty, jak zajišťovat konsistenci dat. Práce je vynikající ukázkou propojení lingvistického a informatického pohledu na velké anotované jazykové korpusy. Nepřehlédnutelný podíl Mgr. Jana Štěpánka na konečné podobě PDT 2.0 je vedlejším produktem práce na této disertaci.

V Praze 24. června 2006

Prof. PhDr. Jarmila Panevová, DrSc.,
školitelka