Vyjádření školitele k dizertační práci

Pavel Straňák: Annotation of Multiword Expressions in the Prague Dependency Treebank

Pavel Straňák se ve své disertační práci zabýval problémem definice, specifikace a anotace víceslovných výrazů na pozadí Pražského závislostního korpusu.

Cíl práce

Cílem práce bylo analyzovat chování víceslovných výrazů (MWE, multiword entities) v reálných textech (korpusu), a to právě a jen jich: tj. i když by v řadě případů bylo možné zároveň studovat i chování jednoslovných terminologicky orientovaných výrazů (včetně např. pojmenovaných entit), autor měl za cíl zůstat u výrazů víceslovných. Tato analýza pak měla formu anotace MWE na materiálu Prašného závislostního korpusu (PZK), který poskytoval (a zejména bude poskytovat do budoucna) materiál vhodný pro další zkoumání MWE ve vztahu k syntaxi povrchové i hloubkové, k čemuž právě PZK poskytuje dostatek rigorózních možností vzhledem k tomu, že anotace MWE v této disertační práci je s ním propojená. Cílem bylo nastínit i některé hypotézy o souvislosti hloubkové syntaktické reprezentace a MWE.

Obsah práce

Práce je rozdělena do osmi kapitol. Po motivační části a krátkém úvodu autor v kapitole o MWE popsal jejich definice v literatuře u autorů, kteří již s tímto pojmem pracovali. Ve třetí kapitole pak následuje stručný „úvod do PZK", se kterým je práce velmi úzce svázána. Od čtvrté do sedmé kapitoly pak autor popisuje vlastní práci: definici MWE v práci použitou a nově vyvinuté anotační schéma kompatibilní s PZK a jeho nástroji (kap. 4), anotační nástroj SemAnn (kap. 5) a strukturu slovníku SemLex (kap. 6), který vznikl extrakcí z anotovaného korpusu, nicméně jeho strukturu (a editační možnosti a nástroje) bylo třeba vypracovat zvlášť. V sedmé kapitole popisuje vlastní anotační proces, nově vypracovanou míru pro anotátorskou shodu na MWE, a rovněž dvě formálně lingvistické hypotézy (o „souvislosti" MWE vzhledem k anotaci na tektogramatické rovině PZK a identické struktuře každé MWE v této reprezentaci). V osmé kapitole autor uzavírá práci a věnuje i značnou pozornost možné budoucí práci na MWE a využití anotovaných dat, která v této práci spolu s anotátory připravil.

Hodnocení

Předložená práce, ačkoli je na první pohled velmi stručná, dokumentuje provedený výzkum a to, že cíle práce byly splněny: byl vytvořen slovník MWE SemLex a k němu anotovaná data (jako přídavná vrstva anotace nad PZK). Tato práce se přitom neobešla bez teoretických úvah a předběžného zkoumání chování MWE v češtině; tyto úvahy a explorace korpusů pak vedly k návrhu specifikace anotace (anotačních pravidel) a vlastní anotaci. Při této práci byl rovněž velice pečlivě rozebrán způsob vyhodnocení anotátorské shody. Lze konstatovat, že i když absence (alespoň v kopii – pravidla existují, ale nebyla do práce včleněna) anotačních pravidel neumožňuje konkrétně zhodnotit v detailech všechny nové aspekty autorova přístupu, tak zpracování dvou zmíněných hypotéz ukazuje, že připravený materiál je velmi cenný. Jako největší přínos pak hodnotím to, že tento slovník byl na rozdíl od řady jiných českých slovníků (byť rozsáhlejších) připraven výhradně na základě anotace dat (po pečlivém zvážení a specifikaci anotačního schématu), tj. „zdola" (podobně jako některé zahraniční slovníky – jmenujme např. Cobuild), a nevzniká tedy otázka (ne)konzistence přístupu a vůbec teoretické definice „slovníkového hesla." Je ovšem škoda, že výsledný slovník (případně jeho frekvenčně

„zajímavá“ část) není součástí vlastní práce (nebo alespoň tištěné přílohy), neboť i ten lze považovat za její součást (slovník je samozřejmě k dispozici elektronicky včetně vizualizačních a vyhledávacích nástrojů v rámci PZK i samostatně).

Závěr

Ačkoli předložená práce ve své písemné podobě vykazuje některé výše uvedené problematické body, považuji ji za práci splňující kritéria disertační práce na MFF UK a doporučuji ji k obhajobě.

[redacted]

Praha, 16. 8. 2010, Jan Hajič, ÚFAL MFF UK