Posudek na diplomovou práci

Marie Konárová: „Školní větné rozbory jako možný zdroj závislostních korpusů (?)“

Předložená práce srovnává anotaci morfologie a závislostních syntaktických struktur češtiny, jak je prováděna na Ústavu formální a aplikované lingvistiky MFF UK (v Pražském závislostním korpusu, PDT) se zvyklostmi větného rozboru na českých základních a středních školách (školní větný rozbor, ŠVR). Práce si klade za cíl zjistit, do jaké míry lze případně využít školní větné rozbory jako ručně anotovaná data analogická Pražskému závislostnímu korpusu.

Práce je členěna do 5 kapitol, Závěru a příloh. Kapitola 2 pečlivě popisuje autorkou využívaná externí data a nástroje, zejména editor větných rozborů Čapek. Kapitola 3 popisuje metodiku a problémy spojené se získáváním dat na školách. Kapitola 4 popisuje transformaci ŠVR do formátu PDT a v kapitole 5 se vyhodnocují výsledky. Významnou část práce tvoří návrh a hodnocení metodiky práce se žáky základních škol. To činí diplomovou práci v rámci oboru informatika poněkud atypickou, avšak určitě ne nezajímavou. Na druhou stranu vlastní informatický i lingvistický přínos tkví zejména ve čtvrté kapitole a v několika stech řádků konverzních skriptů na přiloženém CD. Uvítal bych jasnější deklaraci autorství editoru Čapek, který je sice uveden mezi externími nástroji, avšak není citován žádný zdroj a z pozdějších častých zmínek o jeho vývoji lze podlehnout dojmu, že tento editor je ústředním softwarovým přínosem práce. Zajímalo by mne, zda se autorka přímo podílela na jeho úpravách, nebo zda pouze tlumočila požadavky na úpravy někomu jinému.

Práce je dle mého názoru dobře strukturována, ve většině případů je zřejmé, co autorka dělala a proč. Výsledky jsou prezentovány v řadě tabulek a diskutovány, i když v některých případech by přesto stály za podrobnější rozbor (viz níže). Práce je psána česky, gramatické chyby či překlepy nejsou příliš časté. Použitá literatura je citována na odpovídající úrovni.

Přiložený disk obsahuje autorčiny konverzní skripty v Perlu a dotazy v SQL. Neškodilo by, kdyby obsahoval i editor Čapek, pokud to umožňuje jeho licence. Určitě je ale velká škoda, že nejsou přiloženy anotované soubory získané od žáků, učitelek a zlatý standard od anotátorky PDT.

Závěr výzkumu nevyznívá příliš ve prospěch počáteční hypotézy o využití školních rozborů jako zdroje dat. Přínosné je však, že práce mapuje příčiny tohoto závěru. Domnívám se, že zkušenosti z této práce a z předcházející práce Ondřeje Kučery (STYX) by bylo možné propojit např. ve webové aplikaci, kde by si žáci mohli procvičovat větný rozbor dobrovolně a dlouhodobě, přičemž aplikace by jim zvýraznila chyby.

Konkrétní připomínky a dotazy

- Za jedno z nejslabších míst považuji svázání morfologické analýzy s celým větným členem a ztrátu informace při slučování uzlů. Skutečně nebylo možné navrhnout editor lépe?
- Str. 30, tabulka 4.3: objevují se druhy větných členů („analytické funkce“ ve ŠVR), ale chybí tabulka s vysvětlivkami. Nebo alespoň odkaz do příloh, ale nenašel jsem ji ani v přílohách. Navíc nevím, proč Čapek ukazuje uživateli tyto zkratky a ne celý název alespoň tam, kde je jednoslovný.

- Je škoda, že transformace zahazují informaci o druhu příslovečného určení, když už ji tam učitelky dodaly.
- Str. 13: „interpunkce je přidružena ke slovu, které se nachází před ní“ — takže i třeba levá závorka? A co na začátku věty?
- Str. 19, obr. 2.9: Žáci na základní škole spojují „měl příležitost“ do jednoho větného členu?
- Str. 31, popis pravidel: „přiřazuji nejčetněji se vyskytující odpovídající analytickou funkci“ — Jak často tohle vedlo k chybě?
- Str. 33, pravidla pro určování slovního druhu: druhou pozici značky PDT přece nemáme šanci určit, když nemáme informaci o druhu číslovky, spojky atd. Ale šlo by zapojit automatickou morfologickou analýzu (která bere v úvahu slovní tvar) a anotace žáků použít jen jako disambiguaci.
- Str. 34, tabulka 4.12: Jak to, že slovo „zelený“ má pořadí 1 z celkem 2 tokenů v uzlu, když uzel pokrývá tokeny „byl zelený“ z věty „Mech u cesty byl krásně zelený.“? To pak není divu, že pravidlo selhalo, jak ukazuje obrázek 4.3. Mimochodem, na tomto obrázku je také vidět chyba ve jmenném rodě u slovesa „byl“, protože jmenný rod se v Čapkovi neurčuje, mohli bychom ho ale přece zjistit pomocí stromové struktury a shody.
- Str. 38: „je úspěšnost transformace syntaktické struktury nízká.“ A proč? Postrádám podrobnější rozbor.
- Str. 41, tabulka 5.7: Žáci dosáhli vyšší úspěšnosti u ženského rodu, protože si asi pletou mužský rod se středním. No a jakého rodu je tedy větný člen „s ním“, když nevidím okolní věty dotyčného textu?
- Str. 42, nízká úspěšnost určování stupně: Není to náhodou metodologická chyba, protože u mnoha příslovcí se stupeň neurčuje?
- Str. 42, „Vysoká přesnost určování budoucího času v případě žáků ve škole je pravděpodobně opět způsobena výběrem určovaných slov.“ — Nešlo by to nějak změřit? Např. spočítat, kolikrát se žák vyhnul určení času, když správný čas byl přítomný, budoucí, resp. minulý?

Závěr

Práce jako celek poskytuje zajímavý pohled do oblasti na pomezí informatiky, lingvistiky a pedagogiky. I když použitý postup má ještě rezervy, domnívám se, že práce splňuje požadavky kladené na diplomovou práci magisterského studia a doporučuji ji k obhajobě.

Jenštejn, 22. 1. 2012

RNDr. Daniel Zeman, Ph.D.
ÚFAL MFF UK