Posudek vedoucí práce

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

Cílem diplomové práce Marie Konárové bylo provést studii, zda-li je možné využít školní větné rozbory jako zdroj morfologicky a syntakticky anotovaného korpusu češtiny. Studie pokrývá jak složku datovou, tak složku nástrojů. Anotační schéma cílového korpusu odpovídá koncepci Pražského závislostního korpusu.

Práce je přehledně členěna do šesti kapitol, čtyř příloh, do uživatelské a programátorské příručky a je doprovázena CD. První kapitola je charakteru rešeršního – autorka dohledala několik údajů, jak jsou, či nejsou v jiných zemích větné rozbory začleněny do hodin národních jazyků. Ve druhé kapitole podává přehled externích datových zdrojů a nástrojů, které používá. Třetí kapitola velmi podrobně popisuje sběr dat, který autorka prováděla na vybraných školách. Kapitoly 4 a 5 jsou klíčové při hledání odpovědi na otázku položenou v samotném názvu práce – transformační pravidla pro převod stromové struktury, morfologických tagů a analytických funkcí popisuje čtvrtá kapitola; jejich evaluaci je věnována kapitola pátá. V Závěru autorka s výhradami konstatuje, že školní rozbory jsou možným zdrojem anotovaných korpusu.

Práce je sepsána systematicky, s několika gramatickými chybami v užití interpunkce.

Při celkovém hodnocení předložené práce je třeba vyzdvihnout, že takto navržené téma nebylo dosud zpracováno pro žádný jazyk. Objektivně je třeba dodat, že zejména pro to, že se na procvičování větných rozborů neklade důraz. Začlenění softwarové aplikace do hodin českého jazyka představuje inovativní prvek. Vysoce oceňuji přístup a nadšení, se kterým studentka vystupovala před češtináři a jejich třídami. Materiál a zkušenosti, které posbírala, jsou bezesporu cenné. Návrh, implementace a evaluace transformační procedury v porovnání s pěkně zpracovaným sběrem dat je slabší částí práce. Tento fakt podtrhuje i obsah CD, na kterém postrádám dokumentaci jeho obsahu a datové soubory k otestování transformační procedury. Toto svoje hodnocení doplňuji seznamem poznámek a komentářů:

1. PML je datový formát pro reprezentaci dokumentů a jejich anotací, zatímco anotační schémata Pražského závislostního korpusu jsou systémy značek a datových struktur pro anotace dokumentů na jazykových rovinách. Není tedy správné psát cit. „ ... *je v PML využíván patnáctimístný systém budování tagů* ...“ (viz část 2.3.4).
2. Pojem *token* je definován až v části 4.1, jakkoli je používán i v předchozích kapitolách.
3. Tabulka 4.3 obsahuje syntaktická pravidla pro transformaci uzlů stromové struktury, které neobsahují sloveso *být* nebo zvratné *se*/*si*. Jak se provádí transformace uzlů s *být* a *se*/*si*? Pravidla je vhodné doložit ilustrativními příklady.
4. Tabulka 5.1 shrnuje míru shody jednotlivých morfologických kategorií v rozborech provedených kantorkami. Je vhodné uvést míru shody i pro jednotlivé analytické funkce, ne pouze shodu na všech analytických funkcích souhrnně. Shoda na stromových strukturách není uvedena.

5. Část 5.2.2 shrnuje úspěšnost transformace stromových struktur. Uvedení úspěšnosti 37% bez podrobnější analýzy není vhodné.
6. Ve vysvětlování dosažené úspěšnosti (např. strana 39, první odstavec) se často operuje tím, že věty ve vzorku jsou víceznačné. Toto tvrzení je vhodné doložit příklady.
7. Žáci doma zpracovali všechny věty ze vzorku dat, tj. 101 vět, a žáci ve školách zpracovali pouze několik vět. Obsahují tabulky v části 5.3 údaje na větách, které analyzovali žáci obou skupin? Analýza úspěšnosti ve vytváření stromových struktur není uvedena.
8. Automatická procedura pro větný rozbor, tzv. parser existuje i pro češtinu. Vyhodnocení její úspěšnosti na vzorku '101' je vhodné uvést, zejména pro posouzení kvality transformační procedury s ohledem na položenou otázku.

I přes uvedené výhrady doporučuji práci k obhajobě.

V Praze 16. ledna 2012

Mgr. Barbora Vidová Hladká, PhD.
ÚFAL MFF UK