# Oponentský posudek disertační práce

Václav Novák: Semantic Network
Manual Annotation and its Evaluation

## Obsah práce

Předložená práce zkoumá možnosti rozšíření Pražského závislostního korpusu (**PDT**) o další rovinu, která by zachycovala sémantiku textů a dala by se použít jako báze znalostí pro systémy odpovídání na otázky, získávání informací a podobně. Autor formuluje kritéria, která by sémantická reprezentace měla splňovat, a ve vztahu k nim představuje existující formalismy. Jako nejvhodnější se jeví sémantická síť **MultiNet**, její podrobné zkoumání a použití pro rozšíření **PDT** je hlavním tématem práce.

Autor stručně představuje základní prvky sítě **MultiNet**, některé části specifikace přitom upravuje a přizpůsobuje anotačnímu procesu a **PDT**. Vztah struktur sítě **MultiNet** a tektogramatických stromů je dále podrobně popsán: je uvedeno, kterým funktorům obvykle odpovídají které relace a funkce sítě **MultiNet**, podobně se rozebírají i ostatní atributy tektogramatické roviny. Dále je popsána ruční anotace vybraných vět a problémy, které systému sítě **MultiNet** způsobuje reálný text. Anotovaná data jsou dále velmi podrobně evaluována, zkoumána je hlavně shoda anotátorů. Pro atributy, relace i funkce sítě **MultiNet** jsou uvedeny typické rozdíly v anotacích.

Práce je přehledná, psaná anglicky, srozumitelně, jen s malým počtem chyb. Doplňuje ji rejstřík, devítistránkový seznam použité literatury a přílohy.

## Přednosti práce

Sám autor poznamenává, že projektu **PDT** chybí sémantická rovina. Jeho práce je významným pokusem o její návrh (a tím je práce navýsost aktuální), ale navíc přináší i praktickou zkušenost s jejím vytvářením. Čtvrtá kapitola, podrobně sledující mezianotátorskou shodu, je velmi zajímavá i z hlediska metodologického – hledá totiž odpověď na otázku, jak shodu měřit v případě komplexní strukturní anotace, kdy jsou klasické metody vlastně nepoužitelné. Podrobný rozbor rozdílů v anotaci hodnot atributů a typů relací či funkcí ukazuje nedostatky v jejich specifikaci či hraniční případy, typické pro přirozený jazyk.

Jedním z důležitých výsledků práce je autorovo doplnění některých kritérií, která by měla splňovat sémantická rovina popisu jazyka.

Srovnání tektogramatické roviny a sítě **MultiNet** se může stát zdrojem dalšího rozvoje obou rovin popisu jazyka.

## Problematická místa a otázky

Po formální stránce nelze práci téměř nic vytknout. Několik drobných nepřesností se v ní však vyskytuje, žádnou z nich ovšem nepovažuji za závažnou.

Některé vlastnosti **PDT** jsou popisovány zbytečně zjednodušeně:

- Na str. 19 se u morfologické roviny nezmiňuje přiřazení lemmatu.
- O analytickém stromě autor píše (tamtéž) jako o zjednoznačněné („unambiguous") závislostní reprezentaci. Nezmiňuje se přitom o analytických funkcích `AtrAdv`, `AtrObj`, `AtrAtr`, `AdvAtr` a `ObjAtr`, které lze chápat jako zachycení nejednoznačnosti v závislostním vztahu.
- Pro prvky nižších rovin (tamtéž) používá autor označení „words", i když běžně se používá „tokens", protože mezi ně kromě slov patří i interpunkce.
- Na str. 20 se tvrdí, že na tektogramatické rovině jsou jednotlivé věty od sebe izolovány. To zcela pomíjí použití funktoru `PREC`, který slouží právě k vyznačování mezivětných vztahů, a rozšiřuje tak původní koncepci **FGP**. Navíc mohou být věty propojeny i odkazem na stejný uzel analytické roviny.
- Popis anotace aktuálního členění v **PDT** na str. 36 není zcela přesný: hodnoty atributu `tfa` neoznačují část kontrastivního jádra, jádra a ohniska, ale uzly kontrastivně kontextově zapojené, kontextově zapojené a kontextově nezapojené.

V práci je několik překlepů a ne zcela šťastných či srozumitelných formulací. Jejich počet vzhledem k rozsahu práce je však zanedbatelný:

- Na str. 21: „It is important to be robust..." Pro koho (co) je to důležité?
- V popisu obrázku 2.3 na straně 34 je chybně uveden identifikátor uzlu.
- Na str. 38 u věty (b) mi její možné pokračování v závorce nijak nepomohlo odhalit její smysl.
- Na str. 99 se o dílčích úkolech při anotaci říká, že jsou „not easily separated". Vhodnější by asi bylo „not easy to separate".
- Drobné překlepy: kapitalizace „It seems" na str. 24, chybné číslo strany na str. 33, člen na str. 49 u popisu funktoru `PAT`, chybějící čárka v seznamu na str. 54, velké „G" v názvu kapitoly o slovníku **HaGenLex**, překlep ve slově „frequently" na str. 92.

Obrázek 4.5 na straně 84 vypadá sice hezky, ale jeho srozumitelnost by notně zvýšilo uvedení hlaviček sloupců a řádků (uvedení škály u prvního sloupce a poslední řádky chápu jen jako estetický prvek, žádný graf tam totiž neleží).

Tabulka 4.7 na str. 95 uvádí, jak často se anotátoři neshodli na jednotlivých typech relací. Zajímavé by bylo doplnit tuto tabulku o další sloupec, který by uváděl, kolikrát se daná relace v datech vyskytla (tedy obdobu tabulky 4.5 ze str. 93), a samozřejmě o podíl těchto dvou čísel.

Na straně 55 se uvádí, že ke gramatému `DISPMOD` neexistuje v sémantické síti **MultiNet** protějšek. Významový rozdíl mezi větami „Hokej hraje dobře" a „Hokej se mu hraje dobře" se však jistě nějak vyjádřit dá.

## Závěr

Autor v předložené práci prokázal, že dokáže samostatně tvořivě a vědecky pracovat a řešit složité problémy z oblasti zpracování přirozeného jazyka. Vlastně prokázal ještě víc – schopnost být vedoucím mezinárodního týmu vytvářejícího složitý soubor anotovaných jazykových dat. Doporučuji, aby práce byla přijata jako disertační a aby Matematicko-fyzikální fakulta Univerzity Karlovy v Praze udělila po ukončení disertačního řízení Václavu Novákovi titul Ph.D.

Jan Štěpánek
Karlova Univerzita v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky
10. července 2008