Oponent:
Prof. Dr Patrice POGNAN

Posudek o doktorské disertační práci
pana Jana ŠTĚPÁNKA

Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat)

Jan ŠTĚPÁNEK

Důležitost, ba i nutnost, (co největších) textových korpusů, ať pro účely všeobecné lingvistiky nebo pro řízení různých prací v oboru automatického zpracování jazyků, se již hájit nemusí.

Pražský závislostní korpus (v angličtině „Prague Dependency Treebank", dále „PDT") není jenom jeden z největších korpusů, je také díky vyspělosti a preciznosti svých anotací (ty jsou nutné pro vyhledávání údajů) světovým příkladem.

PDT, který má jako podklad Sgallovu teorii funkčního generativního popisu, se od něj liší hlavně tím, že kromě nulté úrovně (úrovně původních textů) má jenom tři roviny popisu: 1. morfologickou rovinu 2. analytickou rovinu a 3. tektogramatickou rovinu.

Obě poslední roviny jsou tzv. „strukturními" rovinami, což znamená, že reprezentace vět je na nich struktura ve formě stromu, závislostního stromu.

Při konstrukci PDT se používalo hodně automatických postupů, a to:

1. na základě předešlých zkušeností pro analýzu morfologické roviny, analytické roviny a tektogramatické roviny;
2. pro kontrolu anotací, ať byly provedeny automaticky nebo ručně;
3. konečně pro zjištění počtu změn.

Etapy analýzy a anotací přinášejí výjimečně cenné informace o češtině a jejím automatickém zpracování, i když jsou jevy, které nebyly zatím propracovány.

Ale přece na všech úrovních tohoto obrovského díla jsou potíže, s nimiž se musel autor disertace vyrovnávat a navrhnout jejich řešení nebo překonání. Mezi nimi můžeme uvést:

- chyby v původních textech.
- správnost jazyka původních textů, hlavně žurnalistických, nebyla zkontrolována. Jinak by byla neprojektivita struktur značně klesla, pravděpodobně minimálně o polovinu.
- nedůslednost lidské anotace a složitost automatické anotace.
- neadekvátnost lingvistického hlediska (např. složené předložky)
- stromové struktury analytické roviny a tektogramatické roviny si musejí odpovídat, i když mají rozdílné charakteristiky.
- v popisu stromových struktur není všechno závislost, což tedy vyvolává zásadní potíže. Je to vidět, zvlášť na analytické úrovni, u spojovacích a předložkových struktur. (Myslíme, že by se mělo upustit od stromů, které plně nevyhovují, a přejít k jiným řešením, ať to jsou stromy s více dimenzemi nebo struktury ve formě sítí).

V PDT se však dostaly spojky a předložky jako řídící uzly nad významové slovo. Kvůli tomuto zápisu se musely vymyslet pojmy „efektivních" rodičů a synů a „průchozích" uzlů.

Tady také začíná práce pana Štěpánka. Po úvodu a po první kapitole, která má za úkol seznámit čtenáře s funkčním generativním popisem (FGP) pana profesora Sgalla, s pražským závislostním korpusem a s rozdíly mezi nimi, pan Štěpánek uvádí nástroje, které vyrobil pro informatické řešení spojovacích a předložkových konstrukcí. Jedná se tady hlavně o automatické hledání efektivních rodičů a synů. Přitom pan Štěpánek srovnává vlastní metody s cizími metodami, jako jsou např. metody z německého korpusu TIGER nebo Meľčukovy metody „seskupování".

Dále podává pan Štěpánek dlouhý a přesný popis kontrol, které byly rozděleny do tří skupin: *find*, *fix* a *check*. První skupina „find" obsahuje makra, jejichž úkolem je hledání podezřelých jevů, totiž míst, kde může být chyba. Makra ze skupiny „fix" opravují chyby. Makra ze skupiny „check" obsahují příslušné výjimky určitého jevu a kontrolují opakovaně, jestli se v datech při opravách nevyskytují nové chyby.

Z osmi set maker rozdělených do těch tří skupin jich vypracoval sám pan Štěpánek dvě stě čtyřicet čtyři, totiž přes třicet procent celkového počtu.

Účel těchto maker se může rozdělit do dvou skupin: do skupiny technických kontrol a do skupiny lingvistických kontrol na morfologické úrovni, na analytické úrovni a na tektogramatické úrovni.

Pan Štěpánek podává pečlivý obraz systému PDT a přitom přesně definuje svůj podíl na této rozsáhlé a komplexní společné práci. Rovněž pečlivě poukazuje na práce jiných odborníků.

Vysoce odborná a kvalitní práce pana Štěpánka je pro PDT nesmírným přínosem. Ve své disertaci dokázal o ní referovat přesně a skromně.

Doporučuji neprodlenou obhajobu.

V Paříži, dne 28. srpna 2006