Příloha 6: Posudek oponenta habilitační práce

Masarykova univerzita	
Fakulta	Fakulta informatiky
Habilitační obor	Informatika
Uchazeč	RNDr. Aleš Horák, Ph.D.
Pracoviště	Fakulta informatiky
Habilitační práce	Computer Processing of Czech Syntax and Semantics
Oponent	prof. PhDr. Patrice POGNAN, Docteur d'état ès lettres
Pracoviště	INALCO a Université Paris Sorbonne (Paříž)

Posudek na habilitační spis pana RNDr Aleše Horáka, Ph.D.

Uchazeč podal k návrhu na habilitační řízení v oboru „informatika" životopis (40 str.) a monografii s titulem „Computer Processing of Czech Syntax and Semantics" (230 str.).

Název monografie poukazuje na dvě složky: syntax a sémantiku, které jsou pro automatické zpracování jazyků základní. Tato monografie je přitom mnohem bohatší a je složena ze tří hlavních částí, které bych mohl označit jako lexikografie a slovníky, syntaktická a logická analýza češtiny a extrahování sémantiky pomocí logiky. Důležitost první části a důraz, který je na ni nenápadně kladen přímo uchazečem, jsou patrné třeba z krátkého závěru monografie, který obsahuje necelé čtyři stránky, z čehož dvě plné stránky jsou věnovány této první otázce. Poměr mezi těmito složkami je patrný taky z váhy jednotlivých kapitol: druhá kapitola „New Language Resources and Tools" – 61 stran, třetí kapitola „synt – Czech Syntax Analyzer" – 49 stran a čtvrtá kapitola „Transparent Intensional Logic as a Way to Semantics" – 39 stran.

První část monografie představuje lexikální složku a uvádí slovník VerbaLex (typu pražského Vallexu), který je rozsáhlý a podává přesné gramatické údaje o slovesech. Tento slovník hraje důležitou úlohu ve všech úrovních jazykové analýzy. Je spojen s Wordnetem, sloužil jako základ pro Balkanet. Je doprovázen jinými systémy jako jsou DWS (systém pro vytvoření a editaci slovníků „on line"), DEB platforma, která umožňuje použití souhrnu slovníků a která je už implementována na různých místech ve světě a konečně PRALED projekt společně s ÚJČ, který má za úkol vybudovat českou lexikální databázi.

Druhá část monografie je věnována syntaxi. Celý proces syntaktické analýzy je tady založen na formalismu "head-driven", který připomíná LFG a spočívá v několika modulech: systém morfologické analýzy „ajka", slovník "VerbaLex" se slovesným rámcem, systém „Synt" pro analýzu hluboké syntaxe a rozhraní CDW, které má za účel lepší přístup k vyvíjení gramatiky.

Při analýze se pan Horák snažil omezit kombinatorickou explozi použitím 236 metapravidel v gramatice G1 určené odborníkům. G1 generuje G2 a G3, které jsou gramatiky s čím dál tím větším počtem pravidel (respektive 2741 a 11088) a jsou určené automatické analýze. Po aplikaci omezení, která snižují počet stromů, jsou pak vybrány „nejlepší" z nich.

Počet a složitost stromů jsou dále zredukovány různými metodami: filtrovací metodou (zajišťuje kompatibilitu s PDT – sekce 3.3), použitím informací valenčního rámce (sekce 3.4) a filtrovací redukční metodou (sekce 3.5), aby se dostaly tzv. „hezké stromy". Nicméně tyto stromy si udržují složkovou strukturu.

Třetí část monografie je věnována logice jako podklad a prostředek k sémantickému rozboru textů. Autor se opírá o transparentní intenzionální logiku (TIL) P. Tichého, která

sama odkazuje na práce Fregeho a na práce Churche (teorie typů). Tento směr logiky se zdá být schůdným pro lingvistický výzkum. Tato část se věnuje logické analýze vět a je tudíž namířena na slovesný rámec, který je teprve ilustrován příklady. Autor uvádí aplikace a zejména komplexní báze znalostí "Dolphin".

V těchto třech částech jsou určité konstanty, zejména slovník VerbaLex, lexikální nástroj, který se uplatňuje hlavně v syntaktické složce. S ním je úzce spojený slovesný rámec, který se používá nejen u syntaxe, ale taky v logické analýze věty, jak o tom svědčí příklady ze stran 136 až 141.

Všude naráží autor na stejné problémy: složitost složkových struktur, kombinatorickou explozi a v pozadí přítomnost závislostních struktur z pražských i anglosaských pramenů. Z této nesourodosti by mělo vyplývat aspoň uvažování o jiných postupech (viz dotaz č. 2). Uchazeč sice porovnal frázový analyzátor se statistickými závislostními analyzátory, ale aby bylo porovnání správné mělo by být s čistě lingvistickým závislostním analyzátorem.

Z obou dokladů (životopisu a monografie) je patrno, že je pan Horák velice aktivní a angažován ve spoustě výzkumných úkolů a projektů. Je členem šesti současných projektů, z čehož dva jsou evropské. Je řešitelem dvou českých projektů za brněnskou univerzitu. Podílel se na patnácti dalších českých i mezinárodních projektech. Je spoluautorem zhruba sta článků. Je ale třeba podotýkat, že napsal pouze tři články jako samostatný autor, což se může různě interpretovat. Pro mne zrcadlí tento fakt spíš angažovanost pana Horáka a jeho všudypřítomnost ve vědeckém týmu. Zdá se mi velmi důležité, aby někdo spojoval mezi sebou vědecké pracovníky, zajišťoval kompatibilitu a začleněnost rozdílných komponentů výzkumu a dbal na hromadění výsledků. Myslím, že tady je největší zásluha pana Horáka.

Dotazy oponenta k obhajobě habilitační práce

1. Monografie byla publikována v roce 2008 a asi nemalá část odpovídajícího obsahu byla redigována i dříve: „we expect the version Verbalex 2.0 to be available by the end of 2007...". To celkem představuje od 3 do 6 let další práce.

Jaký je aktuální stav platformy DEB II a přípojených nástrojů? Jaké slovníky má veřejnost k dispozici a kolik jich je?

Jak pokročily výzkum a aplikace v oboru automatické analýzy (syntax a sémantika)?

2. Je podivuhodné, že jste jako český badatel používal složkových struktur. Jsou pro takový postup závažné důvody? Uvažoval jste o použití čistě závislostních struktur a aktuálního členění po celém procesu syntaktické a sémantické analýzy? Mně se zdá, že se Tesnierovská a pražská tradice velmi dobře hodí na automatickou analýzu češtiny, která má celou řadu příznivých jevů pro automatickou morfologickou a syntaktickou analýzu oproti angličtině a francouzštině.

3. V lexikální části monografie uvádíte příklady z VerbaLexu. Odkazoval bych hlavně na příklad ze strany 13 „the car bumped to the tree", ale také na dva poslední příklady ze strany 16 a na tabuli na straně 17. Existuje v použití VerbaLexu hranice mezi syntaxí a sémantikou? A když ano, kudy vede?

4. Hlavní heslo ve VerbaLexu je anglické a pochází z WordNetu. Teprve vedlejší hesla jsou česká a jde spíš o překlad nebo aproximaci anglického hesla. Nemá to vliv na pojetí VerbaLexu a nemění to do určité míry popis češtiny?

Závěr

Habilitační práce Aleše Horáka „*Computer Processing of Czech Syntax and Semantics*“ ***splňuje*** požadavky standardně kladené na habilitační práce v oboru Informatika.

V Paříži, dne 16 února 2012 Prof. Dr. Patrice POGNAN