Mgr. Otakar Smrž:
Functional Arabic Morphology
Formal System and Implementation

Oponentský posudek disertační práce

Předložená disertační práce spojuje několik úzce propojených témat při zpracování arabštiny, a to arabskou morfologii (abstraktní model i implementaci včetně vhodné datové základny – slovníku), způsob reprezentace arabské morfologie v rámci systému rodiny Pražských závislostních korpusů a dále řešení technologie zápisu jazyka a jeho transkripce, a představuje tak první příspěvek k formálnímu, leč lingvisticky podloženému popisu arabštiny pro počítačové zpracování.

Práce je členěna na devět kapitol a závěr (a další „povinné" části, jako je literatura, index atd.). O struktuře práce se dozvíme spíše v abstraktu (a obsahu díla), který předchází předmluvu s poděkováními.

V první kapitole (Úvod) autor popisuje dosavadní přístupy k morfologii arabštiny a čtenáři ukazuje, proč tato jeho práce vůbec vznikla – totiž, že současné přístupy k arabštině byly nevhodné buď z hlediska teoretického nebo z hlediska počítačové lingvistiky a jejích současných úkolů. Na konci první kapitoly autor vyjmenovává v deseti bodech jeho originální přínos k vědeckému poznání arabštiny ve smyslu jejího počítačového využití a zpracování.

Druhá kapitola pojednává o možnostech transliterace a vůbec zápisu arabštiny, která se odlišuje od ostatních zejména západoevropských jazyků v tom, že i v tištěné podobě je několik možností zápisu (včetně známého běžného „vynechávání" samohlásek). V současné době jsou však stále ještě potíže i s reprezentací arabštiny v počítačové podobě (to se koneckonců týká i češtiny a prakticky všech dalších jazyků vyjma angličtiny) – i tím se druhá kapitola zabývá a čtenáře informuje tak, aby porozuměl zejména příkladům v dalším textu (které jsou ve většině případů psány arabsky, ale někde je použita transliterace, pokud to bylo účelné při přesných kopiích počítačových dat nebo kódu).

Třetí kapitola je z hlediska návrhu systému arabské morfologie nejpodstatnější: shrnuje teoretická východiska, dosavadní přístupy a zároveň zdůvodňuje výběr pohledu a nové metody a přístupy použité a dále popsané v této práci, a to na podkladu kritiky dosavadních přístupů; autor dokládá, že i dosud asi nejznámější pokus o komplexní řešení arabské morfologie z teoretického hlediska, tzv. Kayův čtyřúrovňový model pro nesouvislé („nonconcatenative") morfologie založený na souhláskovo-samohláskových vzorcích, má nedostatky (nemožnost rozlišení některých případů čistě pomocí těchto

vzorců, nemluvě o problémech s generováním v těchto fonologických n-úrovňových modelech založených na Koskenniemiho disertaci obecně).

Ve čtvrté kapitole autor krátce seznamuje čtenáře s jazykem Haskell, jazykem pro čistě funkční programování, který si zvolil pro implementaci pravidel v jeho systému ElixirFM.

Kapitola pátá, nejrozsáhlejší, popisuje ElixirFM, autorovu implementaci „Funkční arabské morfologie" (autorův termín), návrh slovníku pro tento systém a jeho vztah k anotaci Pražského arabského závislostního korpusu (který autor sám s kolegy z Filozofické fakulty UK vytvořil, resp. byl koordinátorem všech anotačních a specifikačních prací).

Šestá kapitola je jakýmsi „uživatelským návodem v příkladech" k obsluze Haskellovské implementace systému ElixirFM; tato kapitola uzavírá část práce, která pojednává a tvorbě uceleného systému arabské morfologie, což bylo hlavním tématem této disertační práce.

Sedmá kapitola popisuje systém tzv. MorphoTrees, který autor vyvinul pro snazší manuální anotaci právě arabské morfologie v rámci Pražského arabského závislostního korpusu. MorphoTrees řeší anotační problém společný pro řadu jazyků, kdy počet možností, mezi kterými anotátor na morfologické rovině musí volit, je tak velký, že jeho úkol je velmi obtížný a náchylný k mnoha chybám. Přitom se často stává (opět, nejen v arabštině, ale i v češtině), že tyto možnosti lze logicky členit tak, aby anotátor mohl v několika krocích pomocí jednodušších rozhodnutí postupně dospět ke správné volbě. Systém MorphoTrees přitom není omezen na anotaci, ale lze jej pochopitelně použít i na prezentaci morfologické analýzy obecně.

Kapitola osmá je pak velmi důležitou spojnicí mezi arabskou morfologií vyvinutou v rámci této práce a mezi anotací v rámci Pražského arabského závislostního korpusu; jde zejména o vztah mezi morfologií a syntaxí. Poslední kapitola (před krátkým závěrem práce) popisuje implementaci modulu Encode Arabic, jak je definován v kapitole druhé.

Hodnocení:

Ačkoli souhlasím s autorovým vlastním seznamem originálních přínosů práce uvedených v kap. 1.5, za jednoznačně největší klad a přínos práce považuji rigorózně a formálně správně pojatý přístup k arabské morfologii a jeho zachycení ve formě volně přístupného slovníku, spolu s formální specifikací informace ve slovníku obsažené. Tento výsledek sám o sobě by mohl splnit nároky na disertační práci; lze předpokládat, že v práci popsaný a autorem vytvořený slovník bude možno v budoucnu používat na řadu aplikací nyní tolik „populární" arabštiny. Prakticky cenná je i implementace modulu „Encode Arabic". Za velký klad je nutno označit i veřejnou dostupnost vytvořených implementací, ať už jde o slovník nebo další programy.

Vzhledem k propojení práce s prací autora na Pražském arabském závislostním korpusu je však zarážející, že vytvořenou specifikaci autor na tomto korpusu nějak formálně neověřil, alespoň z hlediska anotace (například kvantitativním vyjádřením anotátorské shody, zjištěním zrychlení a/nebo zkvalitnění práce - opět ve smyslu např. shody mezi anotátory - při použití MorphoTrees pro ruční desambiguaci apod). V době „předkorpusové" se pochopitelně taková ověření nevyžadovala a nikdo je v disertaci ani nehledal, ale v dnes je třeba v disertaci tohoto typu takové vyhodnocení provést. Tuto kritiku však nelze chápat tak, že by v práci chyběla standardní evaluace – v tomto typu práce se nejedná o vytvoření jakýchkoli desambiguačních nebo parsovacích nástrojů, a tedy evaluace standardního typu (úspěšnost) je samozřejmě bezpředmětná. Neuvedení těchto kvantitativních údajů pak vrhá nepříjemné světlo i na korpus anotovaný podle těchto pravidel a s tímto slovníkem, neboť není např. jasné, jakou konzistenci lze ve výsledném manuálně anotovaném korpusu očekávat.

Podobně by bylo vhodné uvést údaje o vytvořené implementaci – při použití symbolických interpretovaných jazyků jako ~~i Haskell~~ Haskellu se vždy vtírá otázka, jak dlouho zpracování probíhá, zejména v dnešní době obrovských korpusů a textových dat obecně.

Kapitola o Haskellu je vzhledem k odkazu na internet (haskell.org) „úplná", nicméně pro ulehčení čtení práce jako samostatné publikace by bylo vhodnější tuto kapitolu rozšířit tak, aby pokud možno zahrnovala vše to, co čtenář z tohoto obecně poměrně neznámého jazyka potřebuje při čtení obrázků, tabulek, kódu a příloh.

Práce je psána anglicky, velmi pěkným stylem, i když poněkud „hustě" (stručně). Drobné překlepy (např. str. 90, „combinator"?) lze prominout, nalezl jsem jich jen velmi málo. Za pozitivní lze považovat velmi rozsáhlý soubor citované literatury. Formálně je práce v pořádku (za předpokladu, že český a případně další jinojazyčný abstrakt bude k dispozici v brožuře k obhajobě).

Závěr:

Autor prokázal, že je schopný samostatného vědeckého myšlení a samostatné vědecké práce a doporučuji tedy, aby v případě úspěšné obhajoby a zodpovězení v tomto posudku uvedených nevyřešených otázek mu byl udělen titul Ph.D. v oboru MFF UK I-3 „Matematická lingvistika".

Praha, 15.8.2007