Oponentský posudek disertační práce Mgr. Ondřeje Tichého

Nástroj na tvaroslovnou analýzu staré angličtiny
(Morphological Analyser of Old English)

Posuzovaná disertační práce podrobně popisuje morfologický analyzátor pro mrtvý jazyk: starou flektivní angličtinu, přičemž nezůstává pouze u lingvistické charakteristiky nástroje, ale zabývá se i jeho počítačovou implementací. Lze tedy říci, že je to mimo práci anglistickou vlastně i práce z oboru matematická lingvistika.

Práci tvoří dvanáct částí. Po cílech práce, motivaci, současném stavu poznání ve zkoumané oblasti a volbě vhodné metody se autor pouští do popisu flektivního systému staré angličtiny, probírajе všechny flektivní slovní druhy včetně adverbií (stupňování je pro něho záležitost flexe, nikoli slovotvorby). Ve třetí kapitole se věnuje jednomu z hlavních problémů morfologické analýzy staré angličtiny, obrovské ortografické variabilitě slov, tedy něčemu, co v moderních standardizovaných jazycích neznáme. V tomto smyslu stojí autor před úkolem nesmírně obtížným, pro jehož řešení se obtížně hledají analogie s analýzou moderních jazyků. Ve čtvrté části uvádí po zmínce o morfologických analyzátorech češtiny tři hlavní složky své práce: vstupní data, generátor slovních tvarů a analyzátor využívající vygenerovaných tvarů. V kapitole šesté popisuje zpracování vstupního textu: volbu slovníku, z něhož vychází, nemaje korpus, a gramatiky, přiřazení vzorů slovům podle jednotlivých slovních druhů a příslušná pravidla. Ve větší sedmé kapitole líčí generování slovních tvarů – opět podle slovních druhů – a v kapitole osmé pak analýzu vstupního staroanglického textu. Poté probírá výsledky činnosti analyzátoru: analyzátor je spuštěn na text, který byl manuálně označkován, a autor porovnává výsledky analyzátoru s ručním označkováním textu. V závěru se zamýšlí nad dalším rozvojem analyzátoru, možnostmi zlepšení a uvažuje i o morfologické disambiguaci, pro niž by analyzátor měl představovat dobré východisko. Následuje bibliografie a blok příloh.

Vyjádřím se nyní k některým myšlenkám práce.

V kapitole 1 autor uvádí praktickou motivaci své práce jako pomůcku pro studenty i učitele staré angličtiny. Vědecká motivace se týká korpusového přístupu: u tak speciální odborné činnosti, jakou je výzkum starého jazyka, toho už výzkumník musí hodně vědět o zkoumaném předmětu – zde staré angličtině – a nemůže se tak jako u často adorovaného přístupu corpus-driven spolehnout jen na to, že primárně vyčte něco z dat. Zcela souhlasím: nutnost předběžných lingvistických znalostí je tu pro zkoumání jazykové materie samozřejmostí.

Na s. 13 si autor klade důležitou otázku, zda lze morfologické analyzátory vyvinuté pro standardizované moderní jazyky uzpůsobit pro zpracování jazyků, kde standardizace není/nebyla. Ukazuje se, že je to velmi obtížné, sám jsem si tento aspekt před přečtením této práce dostatečně neuvědomoval, neboť se zabývám morfologickou analýzou a hlavně disambiguací současné češtiny. Stará angličtina je sice typologicky podobná dnešní (i starší) češtině, ale v aspektu variability se stará angličtina od dnešní češtiny zásadně liší. Dále autor velmi pěkně popisuje současný stav poznání ve zvolené oblasti, přičemž vyzdvihuje nijmegenský projekt automatické morfologické analýzy. V pasáži o výběru metodologie je klíčové, že nejde o analýzu jako takovou (rozklad na morfémy), ale o srovnání vstupních tvarů s generovanými tvary na základě slovníku a gramatiky. Hlavní východiska jsou pak uvedena na konci kapitoly 1: „We have no lemmatised or morphologicaly suitably annotated corpus or any set of morphological rules in a digital form beyond digitised texts of standard OE grammars.“ Takže je nutno přijmout metodu založenou na lexikální databázi a pravidlech standardní gramatiky staré angličtiny (s. 18).

V kapitole 2 O. Tichý předvádí morfologickou slo žitost zvoleného jazyka. Od dnešní češtiny se stará angličtina liší mj. tím, že jmenný lexém lze skloňovat podle několika paradigmat, což je v češtině naštěstí výjimečné. Specifickou složitost představují slovesa a různorodost slovesného systému, infixy a změny na morfologických švech, děsivý je počet 132 slovesných paradigmat!

K morfologick ým komplikacím se druží i velká pravopisná variabilita (kapitola 3). Autor si dobře uvědomuje záludnost této variability: chce všechny variantní tvary zahrnout pod totéž lemma (upozorňuje přitom na částečnou homonymii s paradigmatem jiného lemmatu!) a navíc chce zaručit, že jsou to opravdu synonymní varianty mající touž funkci.

V kapitole 4 vytyčuje i na základě znalosti morfologických analyzátorů pro češtinu brněnské provenience svůj plán vytvořit především praktický nástroj. Plán je u vědomí specifické povahy staré angličtiny opodstatněný a dobře zdůvodněný. V dalších částech této kapitoly autor popisuje zvolená slovníková data i problematiku morfologických paradigmat (podle nich pak chce generovat tvary) a na s. 56 specifikuje dva hlavní zdroje své práce: elektronický slovník staré angličtiny a elektronickou gramatiku staré angličtiny s co nejbohatším repertoárem morfologických vzorů. Podrobněji se pak autor věnuje generátoru tvarů na základě slovníku a gramatiky a pak také nejjednodušší složce svého systému: analyzátoru samému.

V kapitole 5 autor dokládá, že se dobře vyzná v informatice,

a v kapitole 6 podrobně popisuje zpracování vstupních dat: probírá slovníky, jež přicházejí v úvahu, a Wrightovu gramatiku staré angličtiny. Podrobně se pak zabývá (kap. 6.2 od s. 68) paradigmaty a pravidly, jak přiřazovat paradigmata tvarům vstupního textu. Předvádí, jak složitá jsou hlavně paradigmata slovesná (s. 71– 73), a v kap. 6.2.2 se věnuje vlastnímu přiřazení paradigmat různých slovních druhů, přičemž komplikace dokládá pro jednotlivé slovní druhy charakteristickými příklady.

V kapitole 7 popisuje autor generování forem, přičemž způsob generování je odlišný pro slovesa a pro ostatní slovní druhy, slovesa jsou přitom rozlišena na silná a slabá. V celé práci a také v kapitole 7 autor předvádí výtečnou znalost morfologie staré angličtiny. Velmi zajímavá jsou čísla na s. 108, která jasně dokládají náročnost vytčeného úkolu: počet 10 796 162, resp. 13 828 441 vygenerovaných tvarů je obrovský a navíc mnoho tvarů jsou gramatická homografa a lexikální homonyma. Čísly na s. 108 „p 109 jsem byl ohromen, ale autor je v následné diskusi (s. 109–112) náležitě zdůvodňuje!

Kapitola 8 se věnuje analýze vstupního staroanglického textu a automatické generaci slovníku z textu včetně porovnávání vstupních tvarů s tvary, jež vygeneroval, a také filtrům pro obtížně zachytitelnou variabilitu slovních tvarů. Existuje mnoho variant různých typů (písařský idiograf, nářeční a diachronní varianty ve staroanglických textech), které se projevují ve variacích nepřízvučných slabik, ve variacích samohláskových i souhláskových a ve variaci záporky. Velmi pěkné a názorné je zobrazení výsledků popisované v odst. 8.2.2. (s. 120) a demonstrované v části 9.1.2.1 *Sample of Automatically Generate Glossary*. Použití pravděpodobnostních filtrů zpracovávajících variaci (s. 122) asi není od věci. Generovaná slova opatřená morfologickými údaji vypadají pěkně a jsou podrobně komentována. Se závěry týkajícími se analýzy lze jen souhlasit.

Výsledná víceznačnost komentovaná na s. 134 a 135 je hrozivá a představuje značný problém pro eventuální disambiguační program. Naprosto souhlasím s úvahami autora o úspěšnosti (či přesnosti, accuracy) z hlediska klasických protichůdných měr: přesnosti (precision) a pokrytí (recall), pokrytí je přitom důležitější. Rád bych při obhajobě diskutoval s autorem o vyhlídkách disambiguačního program, a to v souvislosti s odst. 10.2, s. 148 a 149; tvrdí tam, že Jan Hajič nabízí možný způsob, jak vyvinout disambiguační systém, a to vzhledem k typologické příbuznosti staré angličtiny a dnešní češtiny. Má autor na mysli přístup statistický obecně, nebo nějaký konkrétní rys hajičovského přístupu? Na konci odst. 10.2 spatřuje nejlepší řešení v nějakém statistickém taggeru (na jakých datech se ale

bude učit?). Rozveďme to konkrétněji v diskusi a uvažme i možnost zpracovat data disambiguačními pravidly.

V závěru na s. 149 až 152 autor shrnuje, jak se mu podařilo splnit proponované cíle. Myslím, že s jeho hodnocením mohu souhlasit.

Práce ukazuje, jak je obtížné morfologicky analyzovat starý flektivní jazyk, který se vyznačuje tím, že pro něj dosud neexistuje nějaký větší korpus a že slova vykazují vysokou homonymii i „formální synonymii“ (jedno slovo se zapisuje mnoha variantami). Navíc není jasné, jakou gramatiku zkoumaný jazyk vlastně má: asi by se dala vyvodit jen z jeho dostatečně rozsáhlého korpusu, jelikož jeho mluvčí, kteří by mohli posoudit jeho gramatickou strukturu, už neexistují. Bylo by dobré, kdyby autor neustával v práci v této oblasti a shromáždil nějaký korpus a pokusil se i o disambiguaci jeho textů. Byl by to úkol nadmíru náročný, neboť morfologická analýza je asi vždy snazší (ač velice pracná) než disambiguace analyzovaného textu.

Nejasnosti a formální chyby:

na s. 56, 6. řádek odspoda: Věta *The point...* asi neměla být samostatnou větou, jde o pouhou participiální konstrukci bez finitního slovesa

s. 88, 6. řádek shora: přebytečné *the* před *in*;

s. 97: uprostřed: přebytečné *it* po *linked*;

s. 120: 5. řádek shora: před *simple* asi nemá být *a*.

Pár chybiček jsem mimoto vyznačil ad usum autora přímo v exempláři práce (např. na s. 140, 147).

Závěr

eJe nesporné, předložená disertační práce má velmi dobrou úroveň. Je třeba uvážit obrovskou pracnost autorova výzkumu: autor musel jednak hluboce proniknout do morfologie staré angličtiny, jednak musí umět efektivně programovat. Práce je to tedy zřetelně matematicko-lingvistická, neboť spojuje jazykovědnou fundovanost s informatickou expertízou, a to na velmi dobré úrovni, přitom kvalita v obou oblastech je ze zřejmých důvodů nezbytná. Doktorand tedy odvedl velký kus práce a jasně prokázal své schopnosti samostatně vědecky pracovat. Práci rád doporučuji k obhajobě a doporučuji též, aby doktorandovi byl udělen titul Ph.D. Navíc doporučuji, aby autor v naznačeném směru pokračoval.

V Praze dne 1. 5. 2014

doc. RNDr. Vladimír Petkevič, CSc.
Ústav teoretické a komputační lingvistiky FFUK
oponent