Zápis

z obhajoby disertační práce **Mgr. Ondřeje Tichého**

konané dne 16. 6. 2014

téma práce:

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

přítomní:

prof. PhDr. Aleš Klégr (předseda komise)
prof. PhDr. Jan Čermák, CSc. (vedoucí práce)
prof. PhDr. Libuše Dušková, DrSc.
PhDr. Markéta Malá, Ph.D.
doc. PhDr. Jana Chamonikolasová, Ph.D.
prof. PhDr. Karel Kučera, CSc. (oponent)
doc. RNDr. Vladimír Petkevič, CSc. (oponent)

Předseda komise prof. Klégr zahájil obhajobu a představil kandidáta Mgr. Ondřeje Tichého přítomným členům komise.

Školitel prof. Čermák shrnul doktorandovo dosavadní působení na poli akademickém i pedagogickém a seznámil komisi s tématem jeho dizertační práce.

Mgr. Tichý následně blíže obeznámil přítomné s obsahem a výsledky zpracovávaného projektu. Cílem práce bylo vytvořit počítačový program, jenž by uměl automaticky zanalyzovat jakýkoli staroanglický text, který mu uživatel na vstupu zadá, a na základě této analýzy vypíše seznam všech morfologických funkcí, které jednotlivé slovní tvary v zadaném textu mohou nést. Podnětem pro vytvoření takového programu byla jednak kandidátova zkušenost pedagogická, jednak jeho zájem o využití korpusových metod v diachronním lingvistickém výzkumu. Takovýto program by totiž umožnil:

1) studentům snadno vyhledávat v dostupných slovníkových zdrojích bez předchozí znalosti staroanglické morfologie a ortografických konvencí
2) vyučujícím historického vývoje staroanglického jazyka a literatury snadno kompilovat glosáře k probíraným textům
3) automatickou, respektive poloautomatickou lematizaci staroanglických korpusů.

Svou strukturou předložený program navazuje na obdobné morfologické analyzátory současné češtiny. Příprava a vývoj programu se v zásadě sestávají ze tří základních fází. První je příprava vstupních dat. Lexikální složka pro tento program byla převzata na slovníku staré angličtiny autorů Bosworthe a Tollera (elektronizace tohoto slovníku patří mezi kandidátovy

dřívější projekty), informace o morfologických paradigmatech jsou založeny primárně na Wrightově staroanglické gramatice (1914), ale i na novějších zdrojích (Campbell, 1983; Mitchell a Robinson, 2001; Baker 2012 atp.). Kandidát se v tomto ohledu nedrží tradičního rozdělení paradigmat z hlediska historického, ale uzpůsobuje flektivní vzory pro počítačové zpracování (dochází tedy k určité reorganizaci paradigmat tak, aby vzory s mnoha výjimkami byly rozděleny a rozlišovány zvlášť, zatímco vzory historicky odlišného původu ale se stejnými flektivními tvary byly zpracovávány společně). Na základě těchto vstupních dat je pak ve druhé fázi vygenerován slovník všech flektivních tvarů. Výsledkem třetí, poslední fází, je pak samotný analyzátor. Ten porovnává formy zadané uživatelem s vygenerovanými formami ve své databázi a vyhledává tvary, které jsou shodné. Toto porovnávání probíhá v několika krocích. Program nejprve hledá shodu se slovy uvedenými v konjugačních a deklinačních vzorech, poté s výčty uvedenými ve Wrightově gramatice, a pokud nenalezne hledané slovo ani v jednom z předchozích případů, hledá fonologické a morfologické indicie, na jejichž základě by bylo možno dané slovo k jednomu z existujících paradigmat přiřadit.

Program byl testován na deseti staroanglických textech o celkovém počtu cca 2500 slov. Návratnost se pohybuje v průměru okolo 95% v závislosti na místě a době vzniku textu (návratnost je znatelně vyšší u západosaských textů, jejichž dialekt je v dostupných gramatických příručkách popsán nejpodrobněji). Na závěr kandidát nastínil možnosti dalšího vývoje programu a své plány do budoucna (zlepšení návratnosti, optimalizace vstupních dat, lepší využití morfologických informací a zahrnutí většího počtu dialektových variant).

Po kandidátově vystoupení přednesli oponenti závěry svých posudků. Doc. Petkevič poukázal zejména na obtížnost analýzy staré angličtiny s ohledem na veliké množství introflexe, alografie i homonymie. Zdůraznil, že se předložená disertační práce vyznačuje vysokou náročností projektu a velmi dobrou úrovní zpracování. Prof. Kučera práci ohodnotil rovněž kladně a uvedl, že za jednu z největších výhod programu považuje morfologickou, nikoli morfematickou povahu analýzy (jakou známe např. ze současných synchronních korpusů). Jako klíčový problém vidí riziko přegenerování gramatických tvarů při vytváření databáze paradigmat.

Kandidát Mgr. Tichý se následně vyjádřil k hlavním otázkám vzneseným v oponentských posudcích. Nejprve se věnoval problematice počítačového disambiguování homonymních tvarů zmiňované v posudku doc. Petkeviče. Kandidát uvedl, že pro automatickou disambiguaci lze využít kombinaci přístupu statistického s přístupem založeným na pravidlech. Vezmeme-li při rozpoznávání tvarů v úvahu i okolí zkoumaného slova, můžeme říci, že existují interpretace, které jsou v daném kontextu intuitivně pravděpodobnější než jiné. Částečné disambiguace homonymních tvarů lze proto docílit zavedením pravděpodobnostního principu. Prof. Kučera v této souvislosti vznesl dotaz, zda není možné v tomto případě uvažovat i o počítačem podporované manuální diasmbiguaci. Kandidát uvedl, že s tímto způsobem počítá spíše do budoucna, protože k jeho realizaci je třeba sofistikovanějšího softwaru, než jaký má v současné době k dispozici.

Kandidát se rovněž věnoval otázce přegenerování forem. Zmínil, že řešení tohoto problému má v současné době rozmyšleno jen do určité míry. Generované formy lze například porovnat s tokeny z Torontského korpusu staré angličtiny; tento korpus však nezahrnuje všechny

dostupné varianty všech textů. Vzhledem k povaze dat, které máme u jazyka, jakým je stará angličtina, k dispozici, je navíc nutno počítat s existencí tvarů, které nemáme v dostupných dokumentech doložené.

Diskuse:

V diskusi nejprve vystoupila prof. Dušková s dotazem, jak významnou roli co do počtu identifikovaných forem hrála v kandidátově projektu Wrightova staroanglická gramatika. Mgr. Tichý zdůraznil, že přesná čísla nemá k dispozici, ale lze říci, že počet identifikovaných lexikálních jednotek se bude významně lišit od počtu identifikovaných morfologických forem (v prvém případě půjde pouze o velice malý počet jednotek, množství forem bude naopak vysoké, což dokládá již zmíněná vysoká návratnost v textech, které jsou psány v západosaském nářečí).

Doc. Chamonikolasová se zajímala o zpřístupnění analyzátoru širší odborné veřejnosti, kterou kandidát, jak bylo zmíněno již v úvodu obhajoby, od počátku svého projektu plánuje (s ohledem na bezproblémové zpřístupnění výsledného programu volil Mgr. Tichý při jeho tvorbě pouze běžně dostupný otevřený software). Kandidát potvrdil, že analyzátor bude běžně dostupný, jakmile se podaří doladit některé bezpečnostní a uživatelské záležitosti (jde především o zlepšení algoritmů a dat, optimalizaci uživatelského rozhraní a vyřešení problémů s nároky na počítačový výkon).

Dr. Malá se dotazovala na strukturu kandidátem zmiňovaného Torontského korpusu staré angličtiny. Velikost tohoto korpusu dosahuje více než šesti miliónů slov, Mgr. Tichý nicméně poukázal na to, že korpus zahrnuje i velké množství textů, jež nejsou čistě staroanglické (výrazná část korpusu je tak kupř. tvořena materiálem latinským). Primárním kritériem pro zahrnutí textu do tohoto korpusu je existence kvalitní edice, což ovšem znamená, že přestože korpus v sobě obsahuje téměř veškerý dostupný anglosaský jazykový materiál, od každého textu standardně zahrnuje pouze jednu variantu, čímž pádem neobsahuje veškeré doložené pravopisné i dialektové formy.

Po závěrečné diskusi předseda komise prof. Klégr obhajobu ukončil.

vyhlášení výsledku tajného hlasování:

počet členů komise: 5
přítomno členů komise: 5
kladných hlasů: 5.

Komise navrhla udělit titul doktor (Ph.D.)

Zapsal: Jiřina Popelíková

Podpis předsedy komise: