Posudek na diplomovou práci Bc. Jana Henyše

Registrová variabilita českých internetových textů

Bc. Jan Henyš si pro svoji diplomovou práci nezvolil jednoduché téma: korpusový výzkum registrové variability internetových textů představuje komplikovaný badatelský úkol hned z několika důvodů. Na první pohled zřejmým problémem je šíře a neohraničenost dat, která mají být předmětem výzkumu – internet je stále do značné míry neprobádané území (dosavadní pokusy o zmapování mají povětšinou kvalitativní charakter a omezují se tak na úzkou výsečtů), a není proto snadné vymezit vzorek tak, aby byl věrnou reprezentací zkoumaného fenoménu. Druhým zjevným problémem je absence konvencionalizovaných žánrů či registrů, jimiž by bylo možné se inspirovat a pracovat s nimi podobně, jako se využívají v registrových studiích newebového jazyka.

První úkol vyřešil diplomant tím, že výzkumný vzorek, na kterém si předsevzal registrovou variabilitu zkoumat, odvodil z nejrozsáhlejšího dostupného webového korpusu Araneum Bohemicum Maximum, který je podle všech známých měřítek nejúplnějším obrazem toho, co bychom mohli nazvat český prohledávatelný internet („searchable web“). Neobsahuje tedy copyrightem zatížené texty či texty, pro přístup k nimž je třeba zadat heslo či dokonce zaplatit, není tedy např. věrným obrazem diskusí na sociálních sítích, obsahuje nicméně všechny typy textů, které jsou na internetu *volně* dostupné. To význam práce nijak nesnižuje, tato omezení jsou objektivně daná; naproti tomu práce touto volbou materiálu poskytuje jasnější představu o tom, jaká data můžeme ve webových korpusech očekávat.

S druhým výše zmiňovaným problémem se Bc. Jan Henyš vyrovnává jednak podrobnou rešeršní přípravou, jeho přehled literatury ukazuje na detailní a kritické čtení relevantních děl české i zahraniční provenience, a jednak vlastní invencí. I přes nespornou inspiraci výzkumem D. Bibera a J. Egberta, které lze bez pochyby považovat za vůdčí osobnosti výzkumu v této oblasti, se autor diplomové práce musí se specifickou podobou českých dat vypořádat sám a navrhnout vlastní sadu 22 registrů typických pro český internet. Jeho výběr registrů a jejich definice můžeme považovat za zdařilé. Nezůstává přitom na úrovni formálních náležitostí, ale pokouší se o specifikaci na základě použitých funkčních jazykových rysů (viz tab. 4 na s. 60, kde se mj. ukazuje, proč blog není jedním registrem, ale spíše zastřešující platformou či médiem pro různé typy textů s různými charakteristikami).

K popisu a rozdělení registrů využívá multi-dimenzionální model sestavený pro češtinu v rámci projektu, který aktuálně dobíhá na ČNK. Autor aplikuje výsledky předchozího výzkumu na vlastní data s porozuměním a interpretuje je korektně. Kromě detailních charakteristik každého registru na různých dimenzích modelu poskytuje i vedlejší informace o datech z webového korpusu (např. distribuce registrů či dostupnost textů, což je klíčový parametr pro replikovatelnost výzkumů na webových korpusech).

Bc. Jan Henyš ve své diplomové práci prokazuje nejen to, že je schopen proniknout do teoretických východisek korpusového výzkumu registrové variability, ale zároveň i schopnost exaktní práce s rozsáhlým empirickým materiálem, jeho statistického vyhodnocení a citlivé interpretace výsledků. Výsledná práce tak je teoreticky dobře založená, empiricky odpracovaná a posunující naše znalosti o komunikaci na internetu opět o kus dál. Vzhledem k inovativním závěrům proto doporučuju, aby text byl po přepracování do formy článku časopisecky publikován.

Z výše uvedených důvodů doporučuji diplomovu práci Bc. Jana Henyše k obhajobě a navrhuji hodnocení známkou výborně.

Václav Cvrček

V Praze, 19. srpna 2019