Posudek disertační práce Michala Křena *Diachronní srovnání synchronních korpusů*

Disertační práce Michala Křena se zabývá tématem, které je v současném období vývoje korpusové lingvistiky mimořádně aktuální. Jeho aktuálnost vyplývá ze skutečnosti, že po prvních desetiletích, která byla charakterizována pochopitelnou převládající euforií vyplývající už ze samotné nebývalé možnosti zkoumat jazyky na mimořádně velkých datových souborech, vstoupil jmenovaný obor do další fáze svého rozvoje, v níž si začal postupně uvědomovat svá současná omezení a hledat cesty k jejich minimalizaci, popř. odstranění. Křenova disertace není sice první prací, která se touto problematikou zabývá, ale přesto ji lze – nejen v našem, ale i mezinárodním kontextu – považovat za průkopnickou jak vzhledem k promyšlenosti metody a k hloubce i detailnosti analýzy, tak vzhledem k rozsáhlé podloženosti autorových konkrétních i obecných závěrů.

Cíl předložené práce by bylo možno na nejobecnější rovině formulovat jako zpřesnění dosavadních – zčásti i do značné míry protichůdných – názorů na reprezentativnost korpusů, tj. de facto zpřesnění názorů na vztah mezi korpusy a jazykovou realitou. Konkrétní otázka, kterou si autor v rámci tohoto obecného zaměření klade, je, do jaké míry lze rozdílné kvantitativní charakteristiky získané z rozsáhlých korpusů – navíc z korpusů vybudovaných podle stejného principu reprezentativnosti – považovat za spolehlivý základ pro usuzování o reálných jazykových jevech, o rozdílech mezi nimi a o jejich proměnách. Z možných úhlů pohledu, z nichž by bylo lze tento problém zkoumat, zvolil Michal Křen velmi šťastně diachronní srovnávání synchronních korpusů, při němž důležitost reprezentativnosti vyniká zvlášť výrazně vzhledem k tomu, že vyvozování závěrů o změnách v jazyce na základě korpusů, v nichž jsou realizována různá pojetí reflexe dobových stavů jazyka, je nutně zavádějící.

Materiálovou základnu Křenovy analýzy tvoří korpusy SYN2000, SYN2005 a SYN2010, tedy tři stomilionové korpusy, které byly vybudovány v rámci projektu Český národní korpus s recepcí jakožto hlavním principem reprezentativnosti a které jsou obvykle považovány za nejrozsáhlejší vyvážené vzorky dobových stavů českého jazyka mapující jeho nejnovější vývoj v pětiletých intervalech. Z uvedených tří korpusů sestavil autor promyšleně celkem 48 subkorpusů tak, aby umožňovaly vysoce automatizované, detailní a průkazné zjištění vlivu zastoupení konkrétních textových typů, popř. i jednotlivých titulů, jako je např. Mladá fronta Dnes, na celkovou frekvenci jednotlivých jazykových prvků v korpusu (v dané práci jde především o frekvenci slov a dvoučlenných kolokací). Je třeba ocenit, že s cílem objektivizovat svou srovnávací analýzu těchto korpusů se autor rozhodl pro přístup corpus-driven, tj. přístup pokud možno omezující vliv apriorních názorů, kategorizací nebo očekávání a naopak co nejvíce se přidržující samotných korpusových dat a jejich rozdělení. Vyhraněná snaha po objektivitě ostatně prostupuje celým jeho textem, a to v žádoucí kombinaci se stálým realistickým vědomím možných úskalí, omezenosti či tendenčnosti jednotlivých postupů (viz například minuciózní srovnání výsledků užití různých metod při vyhodnocování zjištěných frekvenčních rozdílů v kapitole 6 nebo komentář k jednotlivým metodám a návrh na jejich vylepšení v oddílech 5.2 a 5.3).

Rozsáhlá analýza, kterou autor realizoval na zmíněných 48 korpusech, je – pokud je možno zvnějšku posoudit – bezchybná, celkově mimořádně pronikavá a inspirující (viz její výsledky spolu s podrobnou mnohostrannou diskusí a množstvím přehledných tabulek a grafů v kapitole 6, s. 82nn.). Mezi četnými – převážně skeptickými – zjištěními týkajícími se možností přímé detekce vývojových změn na základě frekvenčních charakteristik získaných ze synchronních korpusů, vystupuje do popředí především průkazně zdůvodněné konstatování, že synchronní korpusy snažící se reprezentovat „celý jazyk", k nimž patří i

korpusy řady SYN, jsou přes svůj značný rozsah jen málo vhodné k zjišťování „celkových" jazykových změn jdoucích napříč jednotlivými typy textů, a to zejména proto, že není prakticky možné zajistit dostatečně vysokou dlouhodobou kontinuitu skladby korpusů na úrovni jednotlivých žánrů, odborných oblastí ani na úrovni jednotlivých periodik. Vlivu, který mají posuny ve skladbě korpusů na frekvenci jednotlivých slov a dvoučlenných kolokací, věnuje autor ve své disertaci primární pozornost, ale současně na množství konkrétních případů demonstruje i řadu dalších faktorů, které výrazně ovlivňují frekvenční korpusové charakteristiky (za jiné jmenujme například společenské změny, periodické i neočekávané významné události nebo měnící se aktuální témata). Je však potěšitelné, že Michal Křen se nespokojuje s odhalením vlivu tohoto vnějšího šumu, který může budit zdání skutečných frekvenčních posunů v jazyce anebo naopak takové skutečné posuny zcela zastřít. Tam, kde je to možné, navrhuje totiž i konkrétní postupy umožňující některé rušivé vlivy minimalizovat. Kromě toho, že upozorňuje na nezbytnost dokonalejší strukturace, značkování a morfologické analýzy korpusových dat, ukazuje i na nutnost vyvozovat diachronní závěry na základě shody konkrétních frekvenčních charakteristik napříč jednotlivými textovými typy a žánry, navíc s pečlivou následnou analýzou jdoucí v odůvodněných případech až na úroveň jednotlivých titulů.

Předložená disertace Michala Křena vychází z rozsáhlého okruhu české i zahraniční odborné literatury, je zpracována kultivovaným jazykem a vyniká schopností přesně a přístupně formulovat i značně komplikované problémy, postupy a závěry. Jako celek představuje originální přínosný impuls pro teorii i praxi výstavby korpusů a pro korpusovou lingvistiku – impuls, který rozhodně nebude možné v tomto oboru v budoucnosti ignorovat. Jak je zřejmé, jde podle mého názoru o práci, která překračuje běžné požadavky kladené na doktorské disertace, a doporučuji ji tedy bez výhrad jako vhodný podklad k obhajobě i k udělení hodnosti Ph.D.

V Praze 2. května 2012

prof. dr. Karel Kučera, CSc.,
Ústav Českého národního korpusu
Filozofické fakulty UK v Praze

Univerzita Karlova v Praze, Filozofická fakulta
Ústav Českého národního korpusu
nám. Jana Palacha 2, 116 36 Praha 1
tel.: +420 2 21 619 357, ucnk@ff.cuni.cz

Posudek vedoucího vztahující se k rukopisu doktorské disertace

Michala Křena

Diachronní srovnání synchronních korpusů
Diachronic comparison of synchronic corpora

Předkládaná práce Michala Křena je výsledkem mnohaleté praktické zkušenosti s užíváním a výstavbou všech verzí *Českého národního korpusu* i zkušenosti teoretické, promítající se do jeho tvorby. Motivován odborně si zde položil otázku, kterou si dosud klade málokdo a v daném rozsahu materiálu kromě něj nikdo, totiž jak a zda vůbec lze srovnávat korpusy jednoho jazyka ve velmi úzkém časovém rozpětí a v souvislosti s tím i druh přínosu k našemu obecnému poznání jazykových změn za tak krátkou dobu. Autor se důkladně poučil ze všech dostupných a aspoň zčásti relevantních zdrojů z jiných jazyků a dospívá k závěru, že je třeba se vydat vlastní cestou, pochopitelně zásadně korpusovou, nazývanou obvykle corpus-driven a založenou tedy pouze na datech v korpusech nalezených.

Po rozsáhlé analýze problematiky reprezentativnosti a povahy použitelných datových vzorků se rozhoduje pro data jednak v korpusech nejbohatší, tj. publicistiku (s vědomím, že nejde o jazyk celý) a jednak pro způsob jejich studia. Volí v důsledku už osvědčené praxe metodu jejich normalizace umožňující přístup ARF na vzorcích ze třech korpusů časově s pětiletým odstupem; vlastním jeho zájmem a cílem se stávají lexémy a jejich kombinace. Zvláště ten druhý cíl je závažný: protože v tomto v zásadě formálním přístupu nelze dojít k systematickým poznatkům pro jazyk zásadním, totiž vývoji a změnám významu, dostává se k nim aspoň oklikou studiem kombinací (kolokací) lexémů, resp. tvarů ve zvolených vzorcích.

Práce nabízí i v řadě ohledů vhled do postupně rozvíjené metodologie přístupu, kdy se ilustrují starší metody (i domácí), kriticky se upozorňuje na jejich nedostatky a navrhují přístupy vylepšené a modifikované. Týká se to jak srovnávání mezi žánry tak na vzorcích jednotlivých časových údobí uvnitř žánru jediného, vše za použití několika statistických metod, jejich výsledky jsou přehledně pro srovnání tabelizovány s explicitním vytčením, zda jde mezi sledovanými obdobími o pokles v zastoupení formy či její vzestup a tedy vyšší frekvenci. Specificky přitom mj. sleduje systematičnost takového nárůstu či poklesu, což může ukazovat na obecnější tendenci. Rozsah použitých vzorků se omezuje na 50 formálních jednotek. Vhodnost takové metodologie ověřuje do jisté míry i na diachronních korpusech. Obecnou otázkou a problémem však zůstává rozsah dat.

Jakkoliv si je autor plně vědomý úskalí takovéhoto dosud nevyzkoušeného přístupu, je to podle mého názoru přístup a pokus zdařilý a ve volbě lexikonu jako objektu ve srovnání s jinými rovinami zkoumání šťastný. Je nepochybné, že časová blízkost dat ztěžuje možný výklad výsledků (který však nebyl cílem jeho práce, je ale podkladem pro další možné studie), kdy se můžou prolínat skutečná jazyková změna s dobovým výkyvem daným společenskou situací, která se nejrychleji odráží právě v lexikonu, specificky v publicistice, k níž se vztahují jeho hlavní a cenné závěry. Představuje zde, na datech synchronních, vlastně metodu srovnání diachronního, což dobře ukazuje na časově provázanou povahu dat i obou oblastí, synchronie a diachronie.

Práce M. Křena, ve které přesvědčivě zvládl novou problematiku a našel k jejímu uchopení adekvátní formální [illegible] za velmi vítaný příspěvek k teorii tvorby korpusů nabízející i řadu zcela praktických výsledků logie. Velmi rád ji doporučuju k obhajobě i případné budoucí publikaci.

Prof. PhDr. František Čermák, DrSc, vedoucí práce