VÝZKUM CPACT: KOMPUTAČNÍ PSYCHOLINGVISTICKÁ ANALÝZA ČESKÉHO TEXTU

Vladimír Petkevič

Kučera, D., Havigerová, J. M., Haviger, J., Cvrček, V., Komrsková, Z., Lukeš, D., Jelínek, T., Urbánek, T., & Franková, J. (2018). *Výzkum CPACT: Komputační psycholingvistická analýza českého textu*. České Budějovice: Pedagogická fakulta Jihočeské univerzity.

Kolektivní monografie, která je výstupem grantového projektu GAČR, reg. č. GA ČR 16-19087S, popisuje původní výzkum problematiky na pomezí psychologie a lingvistiky, a to na základě rozsáhlých empirických elicitovaných jazykových dat. Tato data v češtině (klíčový aspekt!), zpracovaná metodami počítačové lingvistiky (morfologické značkování dat a multidimenzionální analýza), a jejich vytěžení demonstrují hlubinné souvislosti mluvčího jazyka a jeho psychologických (případně patologických) rysů.

Stěžejní partií monografie je výzkum *CPACT (Computational Psycholinguistic Analysis of Czech Text): Komputační psycholingvistická analýza českého textu* (dále jen CPACT). Obecně lze výzkum v českém prostředí označit za průkopnický, aplikuje totiž metody počítačové (přesněji: korpusové) lingvistiky na české texty v rámci psychologického výzkumu a ukazuje, jak se lidská psychika v osobnostních rysech člověka odráží v jeho jazykových projevech, ať už jako mluvčího či pisatele. Z psychiatrického hlediska umožňuje s větší či menší mírou pravděpodobnosti odhalit psychické patologie: v práci se psycholingvisticky rozebírá deprese jako jedna z hlavních nemocí moderní doby.

Monografie je členěna na šest hlavních částí. Úvodní dvě jsou teoretické: *1. Úvod* zahrnující podkapitolu *1.1 Vymezení základních pojmů*, *2. Komputační lingvistika a psychologický výzkum*. Ústřední kapitola *3. Výzkum CPACT* podrobně popisuje provedený výzkum. Následuje stručný *4. Závěr* s výhledy na návazný výzkum, citovaná (zcela relevantní!) *5. Literatura* a *6. Přílohy*.

V úvodní podkapitole 1.1, nazvané *Vymezení základních pojmů*, je předvedena metoda výzkumu jako metoda komputačně-psycholingvistická. Definují se tu základní pojmy jako *jazyk*, *registr*, *scénář*, *osobnostní charakteristiky*, *psychologický test* pomocí *dotazníku* zkoumající osobnostní charakteristiky za použití různých škál (dotazníkových, testových…)

Kapitola *2. Komputační lingvistika a psychologický výzkum* je členěna do tří podkapitol: *2.1 Současná východiska komputační lingvistiky a její aplikace*, *2.2 Komputační lingvistika v kontextu psychologického výzkumu: Současná témata a vybrané přístupy* a *2.3 Problematika dotazníkového sebeposouzení a posouzení druhým v psychologickém výzkumu*.

Kratičká podkapitola *2.1 Současná východiska komputační lingvistiky a její aplikace* velmi stručně, leč koncizně seznamuje čtenáře s oborem komputační lingvistika, jejími metodami a hlavními aplikacemi/podobory. Podkapitola *2.2 Komputační lingvistika v kontextu psychologického výzkumu: Současná témata a vybrané přístupy* nejprve v krátkosti charakterizuje pojem *psycholingvistika* a speciálně pak *komputační psycho-*

lingvistika. Dále popisuje možnosti zpracování textů pro potřeby psychologického výzkumu s důležitým rozlišením tzv. *uzavřených přístupů* (přístup zvolený autory) a *otevřených přístupů* kvantitativní jazykové analýzy. Jsou rovněž uvedeny významné aplikace v této oblasti. Autoři dále uvádějí psychologicky relevantní osobnostní charakteristiky zjištěné na základě komputačnělingvistické analýzy (zmiňují se např. o klíčové roli funkčních slov či synsémantik), a to na bázi zejména anglofonních výzkumů, ale upozorňují rovněž na výzkumy realizované na jiných jazycích, mj. slovanských. V češtině něco takového dosud chybělo a tento nedostatek monografie více než napravuje.

Podkapitola *2.3 Problematika dotazníkového sebeposouzení a posouzení druhým v psychologickém výzkumu* charakterizuje využití dotazníků v psychologickém testování, jejich meze a související problémy. Zaměřuje se na testování osobnostních charakteristik na základě sebeposouzení (*self-report*) a posuzování druhou osobou (*other-report*). Probírá přítomnost/míru asymetrie mezi těmito variantami posuzování a upozorňuje mj. na možné zkreslení výsledků různými činiteli (sociální desirabilita, problematická interpretace konstruktů a termínů respondentem, specifika samotné introspekce/extraspekce, vztah posuzovatele k posuzovanému aj.). Podává přehled zahraničních psychologických modelů vztahujících se k testování a zasazuje tak výzkum CPACT a v něm použité testování do celkového kontextu.

Hlavní část monografie tvoří její 3. kapitola: *Výzkum CPACT*. Tento výzkum je velmi výstižně popsán ve čtyřech podkapitolách: *3.1 Úvod do výzkumu CPACT*, *3.2 Výzkumný soubor a metody*, *3.3 Charakteristika dat a práce s daty*, *3.4 Zpracování dat a výsledky výzkumu CPACT*; podkapitoly 3.3 a 3.4 jsou dále jemněji členěny. V popisu se jasně zračí interdisciplinarita výzkumu: vedle psychologicky zaměřených podkapitol jsou v kapitole 3 obsaženy i kapitoly počítačovělingvistické a obšírně pojednána aplikace počítačovělingvistických metod na jazykové projevy respondentů. Metodologicky tedy výzkum náleží do oboru psycholingvistika a využívá počítačové analýzy přirozeného jazyka, který se parolově odráží v textech respondentů. Na metodě je ovšem osobité, že spojení psychologického výzkumu a komputačních jazykovědných metod je v českém prostředí ojedinělé: autoři nejen propojili psychologii a komputačnělingvistickou analýzu textu, ale přišli i s vlastní formou testování (úpravou testovacích baterií), přičemž vycházeli z *českých* textů/testů.

Projekt CPACT má dva hlavní cíle:

(i) testovat v češtině hypotézy, jež vycházejí z výzkumů provedených na angličtině a předpokládají vztah mezi psychologicky relevantními charakteristikami osob a formálními textovými parametry komunikátů těchto osob;

(ii) analýzou jazykových dat zjišťovat další proměnné a vztahy, jež dosud nebyly v zahraničních studiích publikovány.

Je přitom jasné, že metody použité v zahraničních studiích nelze jen tak beze všeho převzít a jednoduše aplikovat na češtinu. Autoři tedy vypracovali pět pilotních studií, klíčových pro experimentální psychologický výzkum.

Velmi vhodně byly stanoveny tři moduly: P200 (dvě stě duševně zdravých rodilých mluvčích češtiny reprezentujících obyvatelstvo České republiky z hlediska věku, pohlaví a vzdělání), P20+ (podle plánu dvacet depresivních/úzkostných rodilých mluvčích češtiny, ve skutečnosti jich bylo shromážděno dvaasedmdesát (sic!)) a P2 (dva

posuzovatelé produkovaných textů). V této souvislosti byly formulovány další cíle výzkumu:

(iii) popsat analýzou vztahů (jejími autory byli dva vyškolení posuzovatelé) mezi jednotlivými proměnnými užívání textových parametrů u depresivních/úzkostných pacientů a zjistit odlišnosti od zdravé populace (výzkum se zaměřil na deskripci, exploraci a verifikaci);

(iv) porovnat výsledky škálování s výsledky psychologických testů a s výsledky počítačovělingvistické analýzy (hlavně deskripce). Podkapitola 3.1 popisuje také harmonogram a průběh výzkumu, elektronické rozhraní výzkumu a obsahuje soupis výzkumných pracovníků podílejících se na projektu.

Podkapitola 3.2 detailně charakterizuje výzkumný modul P200 a výzkumný postup. Textový materiál tvořily dva řízené rozhovory a dva psané dopisy lišící se opozicí (a) formálnost vs. neformálnost a (b) dominance vs. submisivita. Úkolem respondentů bylo:

(I) sepsat 1. motivační dopis (žádost o přijetí do zaměstnání), 2. dopis z dovolené, 3. stížnost, 4. omluvný dopis,
(II) 1. absolvovat přijímací pohovor, 2. vyprávět o příjemném zážitku.

Pro psychologické testování byla sestavena původní testová baterie zahrnující 361 položek vytvořených na základě 11 psychologických testů. Příprava a administrace testů jsou podrobně popsány a ilustrovány tabulkami a obrázky, někdy bohužel (vzhledem ke zmenšení) nečitelnými.

Podkapitola *3.3 Charakteristika dat a práce s daty* obsahuje psychometrické charakteristiky použitých testů (3.3.1) a tabulku s podrobným popisem škál testů. Rovněž popisuje osobnostní inventáře (BFI-44: Big Five Inventory a PSSI: Persönlichkeits--Stil- und Störungs-Inventar), dále dotazník úzkosti a úzkostlivosti (STAI X-2: State--Trait Anxiety Inventory), Multimotivační mřížku (Multi-Motive Grid) a různé škály a dotazníky.

V podkapitole *3.3.2 Lingvistická analýza textů* je poskytnut přehled sledovaných textových (zvláště gramatických) kategorií a podán výklad o jednotlivých jazykových parametrech (gramatických, lexikálních) a frazémech a rovněž o kombinovaných textových parametrech. Rovněž je popsána automatická lingvistická analýza textů (větná segmentace, tokenizace, morfologická analýza a morfologická disambiguace), anotace frazémů a další druhy zpracování vstupních textů. Velice důležité a nesmírně zajímavé je na konci podkapitoly srovnání vybraných gramatických kategorií v textech respondentů s obecnou situací v jazyce, reprezentovaném korpusem psaného jazyka SYN2005 (zkoumaného prizmatem žánrů beletrie, publicistiky a odborné literatury) a korpusem mluveného jazyka ORAL; oba korpusy jsou součástmi projektu Český národní korpus. Podíl slovních druhů v elicitovaných textech se, jak bylo zjištěno, nejvíce blíží subkorpusu *beletrie*; texty od respondentů mají vyšší dějovost a spontánnost (vykazují větší zastoupení sloves, zato menší zastoupení substantiv a adjektiv). Byly porovnávány rovněž pády substantiv: opět jsou produkované texty nejbližší subkorpusu *beletrie* a vyšším zastoupením akuzativu a vokativu se blíží korpusu ORAL.

Podkapitola 3.4 se věnuje zpracování dat a výsledkům výzkumu CPACT. Podkapitola *3.4.1 Statistická explorace vztahů mezi osobnostními charakteristikami a textovými parametry* popisuje korekce nesprávně pozitivních (false positives) výsledků jednotlivých provedených testů (Šidákova korekce) a další druhy korekcí. Rovněž uvádí výsledky korelační analýzy lingvistických parametrů šesti produkovaných textů vzhledem k psychologickým testům a jejich škálám. Ukazuje se, že má smysl hledat vztahy mezi textovými parametry a osobnostními charakteristikami, neboť během výzkumu bylo zjištěno celkem 148 prokazatelných vztahů. V podkapitole *3.4.2 Zpracování výsledků dotazníkových variant sebeposouzení a posouzení druhou osobou* se autoři zaměřují na asymetrii výsledků zmíněných dvou variant dotazníkového posouzení. Zkoumá se míra shody v posouzení určitých škál z hlediska jejich pozorovatelnosti a evaluativnosti a též míra shody v posouzení s ohledem na délku vztahu posuzovaného a posuzujícího. Při zpracování a interpretaci dat u těchto dvou variant testování se uplatnily různé statistické procedury, přičemž výsledky ukázaly, že nejvyšší shodu mezi oběma typy posouzení vykazuje *emoční citlivost* jako významná osobnostní charakteristika. Zajímavé je, že škály *neuroticismus* a *přecitlivělost na výrazné chování druhých* vykazují signifikantní odlišnost mezi oběma variantami posouzení.

V podkapitole *3.4.3 Rozsah registrové variability textů* je podrobně popsán multidimenzionální model registrové variability a porovnány elicitované texty projektu CPACT s korpusem Koditex, obecně reprezentujícím variabilitu mluvených a psaných textů. Tento korpus je podrobně charakterizován a rovněž je detailně popsána multidimenzionální analýza (multidimensional analysis — MDA) prizmatem hlavních dimenzí:

1. Dynamický extrém (+) vs. statický extrém (–)
2. Spontánní (+) vs. připravený (–)
3. Vyšší (+) vs. nižší (–) stupeň koheze
4. Polytematický (+) vs. monotematický (–)
5. Vyšší (+) vs. nižší (–) míra explicitní adresnosti
6. Obecný (+) vs. konkrétní (–)
7. Prospektivní (+) vs. retrospektivní (–)
8. Postojovost (+) vs. faktuálnost (–)

Na bázi těchto dimenzí je model MDA porovnán s respondentskými daty podle různých typů testů/textů (motivační dopis, dopis z dovolené, stížnost, dopis s omluvou, motivační pohovor, vyprávění o příjemném zážitku). Ukázalo se, že elicitované texty projektu CPACT nabývají očekávaných charakteristik (respondenti se ve svých odpovědích přibližovali požadovanému registru, vymezenému analýzou MDA, resp. snažili se vyhovět komunikační situaci).

Podkapitola *3.4.4 Odraz osobnosti v textu — rysy modelu Big Five* zkoumá souvislosti mezi osobnostními psychologickými rysy (vycházejícími z modelu Big Five, klasického pětidimenzionálního modelu osobnosti: neuroticismus (emocionální stabilita), extraverze, otevřenost zkušenosti, přívětivost, svědomitost) a morfosyntaktickými i sémantickými charakteristikami češtiny dospělého obyvatelstva reprezentovaného texty od participantů výzkumu z kvótního souboru P200. Analýza celkem 1200 textů přinesla mimořádně zajímavé výsledky:

(i) čím vyšší je extraverze, tím vyšší je podíl zájmen (obecně, nejen například 1. osoby) a příčestí minulého v mluvené řeči (míněno patrně v aktivu);
(ii) čím vyšší je neuroticismus, tím více se vyskytují zájmena v 1. osobě a tím méně neurčité číslovky (např. *několik*) a tím chudší je slovní zásoba;
(iii) čím vyšší je otevřenost, tím bohatší je slovní zásoba, a naopak klesá podíl interpunkce a počet emočních slov a frazémů, u nichž nelze zjistit emoční příznak;
(iv) svědomití lidé málo užívají hovorových a obecněčeských tvarů a také tvarů negativních; naopak ve větší míře užívají substantiv, adjektiv, neurčitých číslovek a posesivních zájmen.

Roste také deskriptivita textu a prokázal se i těsný vztah mezi svědomitostí a výskytem 2. osoby. Ve výzkumu bylo zjištěno 35 spolehlivých korelací, které se v práci detailně probírají. Diskuse je mimořádně zajímavá a interpretace průkopnická patrně i ve světovém měřítku. Dosažené výsledky jsou koncizně shrnuty na konci podkapitoly 3.4.4.4. Bylo zejména zjištěno, že: vztahy mezi osobností a textem jsou obecnějazykové, tj. do značné míry nezávislé na konkrétních jazycích; osobnostní rysy se více promítají do textových charakteristik v komunikačních situacích s neformálním rámcem a do charakteristik mluvených textů (oproti textům psaným).

Podkapitola *3.4.5 Odraz osobnosti v textu — projevy deprese* obsahuje výsledky zkoumání respondentů v klinickém modulu P20+; jde konkrétně o manifestace aktuální deprese v psaném textu. Tato manifestace je ovlivněna především pohlavím respondenta a také jazykovým registrem spjatým s danou komunikační situací. Jak vyplývá z výzkumu, klíčovými a přitom spolehlivými příznaky deprese jsou u respondentů:

(i) nadměrné užívání emočně nabitých slov (platí obecně, nejen v češtině);
(ii) nadměrné užívání osobních zájmen (v singuláru i plurálu), neurčitých číslovek a frazémů;
(iii) menší užívání přítomných a budoucích slovesných tvarů a aktivních l-ových příčestí i — patrně — vyšší užívání pasivních konstrukcí v minulém čase;
(iv) menší užívání vokativu a druhé osoby;
(v) s rostoucí mírou aktuálně prožívané deprese klesá konkrétnost a specifičnost textu.

Jsou to cenné údaje, neboť na základě takovýchto korelací se dají vytvářet prediktivní modely, umožňující identifikovat osoby ohrožené depresí. Opravdu pozoruhodné a obtížně předpokládatelné je zjištění, že nejsilnější prediktivní model deprese pro muže vychází z formálního řečového registru textu stížnosti, zato pro ženy z neformálního řečového registru textu dopisu z dovolené.

Interdisciplinární výzkum uplatněný v projektu CPACT pokládám za přímo modelový příklad ideální mezioborové kooperace. V projektu se výtečně snoubí vynikající expertiza psychologická a lingvistická, která vedla — nepřekvapivě — k netriviálním a průkopnickým výsledkům, daným zejména šťastně zvolenou metodologií, jejíž opodstatněnost se prokázala nade vši pochybnost. Z hlediska lingvistického je velmi potěšitelné sledovat, jak přínosná je jazykovědná multidimenzionální analýza nejen pro zkoumání jazyka samého, ale i z hlediska tzv. vnější lingvistiky, tj. z hle-

diska přesahu k jiným oborům, zde k psychologii. Rovněž jsem velice potěšen tím, jak plodně se uplatňují výsledky netriviálního gramatického značkování českých textů.

Teoretické výsledky představují velmi dobrou základnu, na níž lze v dalším výzkumu plodně stavět. Tyto výsledky navíc nezůstanou, jak pevně doufám, pouze pokladnicí teorie, ale budou využity i prakticky, zvláště zřejmě v psychodiagnostice.

Je velmi sympatické, že autoři nikterak nezastírali obtížnost svého výzkumu, jež byla dána mj. složitou organizací a logistikou (mnoho respondentů, náročná elicitace různých typů textů…).

Práce je průkopnická i ve světovém měřítku. Přesvědčivě prokázala, jak je prospěšné jazykově zkoumat osobnostní rysy člověka nejen prostřednictvím angličtiny, ale i jazyka z jiné jazykové rodiny, který disponuje jinými gramatickými kategoriemi.

Z hlediska jazykového stylu se text dobře čte, formulace jsou srozumitelné a věcné bez pleonasmů a redundancí. Text obsahuje mnoho odborných výrazů převzatých z angličtiny, pro něž patrně (bohužel!) neexistuje český ekvivalent. V psychologicky zaměřených studiích autoři někdy používají po mém soudu zbytečných anglicismů, které asi nejsou součástí odborného diskursu, např. ***facilitovat*** *dostupnost* (s. 24), ***augmentovat*** *nepodstatné detaily* (s. 38). V textech se občas objeví překlepy: *tex**o**vé markery*; *(navenek vyjádřené) chov**a**ní* (oboje na s. 13), ***jisota*** (s. 25); morfologicky spíše hyperkorektní varianty: *jazykových jevů…, které nejsou a priori definované a vychá**z**í* (místo *vycházejí*); nevhodně opisné stupňování: ***více skryté*** *charakteristiky* (s. 39); anglická slovesná valence: *přístupy umožňují lépe* ***diskutovat*** *základní teorie* (s. 25); neobratné formulace: *a dále* ***skutečnost, že to, že*** *je někdo vhodným posuzovatelem předložené situace…* (s. 35); ***bylo pracováno*** *se všemi pěti škálami* (s. 92)…; lexikální a pravopisné chyby: ***Augustiánské*** *a* ***D****escart**e**ovské tradice* a další. Tyto nedostatky však příliš nebrání plynulému čtení, neboť nejsou četné.

Na monografii se podíleli výše uvedení odborníci z různých oborů: psychologové, počítačoví lingvisté, statistikové, práce je tedy výsostně interdisciplinární, příklad interdisciplinární kooperace je tu čítankový. Velkým přínosem textu, jemuž nelze odborně nic vytknout, je logické členění jednotlivých studií (úvody, cíle, hypotézy, závěry se shrnutými výsledky), text se i proto snadno a plynule čte. Integrální součástí textů jsou obrázky, grafy, tabulky mj. s údaji o statistických šetřeních, s přehledy testovacích baterií ad.

Recenzovaná monografie, kterou vydala Pedagogická fakulta Jihočeské univerzity v Českých Budějovicích v roce 2018, zásadně přispívá k modernímu psychologickému i lingvistickému výzkumu. Je názorným příkladem plodné mezioborové spolupráce a rovněž velkou inspirací pro další psycholingvistický výzkum.

V Praze 7. prosince 2019

Vladimír Petkevič | Ústav teoretické a komputační lingvistiky FFUK
<vladimir.petkevic@ff.cuni.cz>