DeníkN

Na začátku spisovné češtiny byl dvě stě let mrtvý jazyk, říká lingvista, který jde proti proudu. Nová Pravidla už možná nebudou

„Brusičství jazyka je náš intelektuální export.“ Lingvista Václav Cvrček. Foto: Gabriel Kuchta, Deník N

„Náš spisovný jazyk vychází z předbělohorské češtiny, o živém jazyce totiž obrozenci soudili, že je v úpadku,“ říká jazykovědec, zástupce ředitele Ústavu Českého národního korpusu Václav Cvrček. A tak podle něj kodifikovaná čeština kulhá míle za tou živou dodnes. Proč by měli lingvisté jazyk jen popisovat a necpat ho do norem? Co je jazykové safari? Mohla by lingvistická analýza pomoci rozbít trollí farmy? A jak přispěla ke zrušení trestu smrti v Británii?

V pořadu Hyde Park v České televizi ve vás lidé chtěli najít jazykového soudce, očekávali, že je utvrdíte v jejich pohoršení nad „nešvary“ s používáním slov, jako jsou *nicméně, potažmo*, nebo že jim dáte za pravdu, že je čeština se svými *červeňoučkými*

kulaťoučkými jablíčky **nejkrásnější a nemá se prznit třeba přejímáním anglicismů, nebo jak je hrozné, že se i v písemném projevu běžně objevuje obecná čeština. Ale to si nevyhlídli toho pravého, že?**

To tedy ne. Já jim na jednu stranu rozumím, ale opravdu tu nejsem od toho, abych rozsuzoval, jaké tvary jsou správné a které ne. Věčný problém. Lidi mají pocit, že k jazyku se má přistupovat normativně, jde o hluboce zakořeněný a asi přirozený pocit, ale nemá moc podporu v tom, co o jazyce víme. Protože normativita… À propos, měli jste teď o tom moc hezkej sloupek.

Myslíte Jste normální? od kolegy Moláčka?

Ano. Ptá se, co je vlastně „normální“, jestli je normalita něco, co je časté, nebo něco, co někdo určil, že je normální. V lingvistice je to podobné. V jazyce jako bytostně demokratické sociální instituci, kterou svým jednáním ovlivňuje každý, podle mého přesvědčení platí, že „správně“ je to, co je časté, protože jazyk je především záležitost zvyku. V jazykové praxi pořád vyjednáváme: o významech, o tom, jaké slovo v jaké situaci využít, jakým stylem mluvit nebo psát při jaké příležitosti, a co je časté, bývá i vhodné. Tak to funguje.

Ale existuje mnohem větší skupina lingvistů a ještě větší skupina uživatelů, kteří si myslí opak: „Ne! I kdyby to bylo jakkoliv časté, máme mít možnost o něčem říct: ‚To je špatně, protože…‘“ A dosaďte si důvod.

Dvě stě let pozadu už na začátku

Podle vás jazyk nepotřebuje pravidla pravopisu, nepotřebuje normy, protože si jde stejně svou cestou a všechny příručky jen klopýtají za živým jazykem. Je to náročná pozice, jít ve vlastním oboru proti hlavnímu proudu?

Někdy ten protiproud cítím silně, ale snažím si udržet integritu, nemůžu psát to, co si nemyslím, takže vlastně nemám na výběr.

S tímto přístupem k jazyku jste se dal na lingvistiku, nebo vaše názory formovalo až to, co jste se o jazyce dozvídal během studia?

Jsem z učitelské rodiny, chtěl jsem být středoškolský pedagog. K profesi češtináře patří, že předává i informace o normě a kodifikaci, ale abych to mohl studentům zprostředkovávat věrohodně, musím si za tím, co říkám, když je nutím, aby něco sami dodržovali, sám stát.

A tak jsem pořád pátral, kde jsou kořeny toho, proč jsou některé tvary „správné“ a jiné „nesprávné“. Proč nesmíme psát „mladej“, ale „mladý“. Ano, tím jsem prošel už v prváku na vysoké, šel jsem k pramenům a zkoumal, jak se utvářely názory na český jazyk od 19. století.

A na co jste přišel?

Že mnohem větší roli v české lingvistice hrál spíš nacionalismus nebo elitářství a vůbec „ušlechtilé“ představy o tom, jak by měl jazyk vypadat a fungovat, než výzkum přirozeného jazyka jako takový. Už od 30. let měl Pražský lingvistický kroužek takovou avantgardní představu, že jazyk je možno řídit, sám Vilém Mathesius napsal, že „*dnešní situace linguistické teorie nám umožňuje, abychom vědeckým zasahováním proces tříbení trochu urychlili“*.

Naplno se to pak projevilo v 50. letech, kdy se do toho promítl marxismus a jazyk byl předmětem centrálního řízení jako vše ostatní. Takový přístup k jazyku ale není vědecký, v regulaci českého jazyka hrály roli jiné věci než živý, opravdový jazyk.

Boj o národní sebevědomí, politika, historická traumata…

Všechno se tam promítá. První generace obrozenců se snaží čistě jen o to, vyrovnat se němčině, ale postupem času se přidává víc a víc nacionalismu. Lingvisti se dodneška přou, jestli měla v 19. století ta první generace jinou možnost než vzít tehdy už dvě stě let mrtvý předbělohorský jazyk a postavit na něm základ spisovné češtiny svých současníků.

A měla?

Určitě ano, ale byla v zajetí tehdejších nálad. O živém jazyce se tehdy soudilo, že je v úpadku, i když ten úpadek nebyl definován na základě lingvistiky.

„Pokud se nikdo neodváží tu bláznivou příručku vydat znovu, přestane být použitelná.“ Foto: Gabriel Kuchta, Deník N

Ale prostě proto, že po bitvě na Bílé hoře bylo „všechno špatně“?

Spousta obrozenců vnímala Bílou horu, a zejména období, které přišlo po ní, jako národní pohromu a převažovala představa, že když se nám nedaří politicky, nemůže se dařit ani jazyku.

A tak si obrozenci řekli, že lepší bude vrátit jazyk do stavu „před úpadkem“?

Do oněch „krásných časů“… A vzali si za základ dvě stě let starý jazyk. Nepřijali tedy ani spoustu náležitých hláskových či morfologických změn – *mladý/mladej, mlíko/mléko*.

Jeden příklad za všechny: první generace obrozeneckých jazykovědců v čele s Dobrovským říkala, že by pro přirozenost a mluvnost jazyka bylo lepší u přechodníků od systému tří tvarů

ustoupit (např. nesa, nesouc, nesouce), protože už jsou dávno mrtvé, a navrhovala ponechat pouze jeden tvar (nesouc) ve všech funkcích. Nicméně Jungmann, o půl generace mladší, řekl, že ne, že tam vrátí i tyhlety drobnosti, protože je to dokladem bohatosti morfologie. Jenže co se stalo?

Přechodníky skoro nikdo nepoužívá, protože je to moc složité.

Přitom kdyby dali na Dobrovského a kdybychom měli jen jeden tvar, pak by byl možná i náš živý jazyk o přechodníky bohatší. Tohle se táhne se spisovnou češtinou od 19. století až do dneška.

Nespisovná (obecná) čeština je důsledek kontinuálního vývoje, zatímco moderní spisovná čeština byla od začátku postavena na diskontinuitě. Vždycky bude umělá a nepřirozená a naše tříbení jí v tomto ohledu dělá spíš medvědí službu. Důsledkem je, že ve škole se ji musíme znovu učit.

Máte školou povinné děti?

Dcera je v první třídě, nedávno mě napomenula, když jsem řekl *vokno*, protože *vokno* „se neříká“. „Ale říká, vždyť já to teď řekl,“ povídám. „Ale není to správně!“ Získala ve škole tuhle informaci, místo aby se dozvěděla, že v obecné češtině je zcela běžné říct *vokno*, ale ve škole v hodině, tedy v nějaké formální situaci, je vhodnější říkat *okno*. Namísto nějaké škály, kdy je vhodné používat jednu nebo druhou variantu, je celá věc prezentována jako černobílá, jako bych udělal dopravní přestupek.

Syn je ve třetí třídě, kde se právě učí vyjmenovaná slova. S hrůzou si uvědomuju, jak moc náš normativní přístup k jazyku, který pramení mimo jiné i z rozhodnutí obrozenců v roce osmnáct set něco, způsobuje, že se moje dítě bude dva roky v jednom předmětu věnovat nácviku něčeho, čemu by se vůbec věnovat nemuselo.

Neintuitivní psaní y/i ve vyjmenovaných slovech je proto, že to „správné“ y/i v těch slovech už neslyšíme, zatímco před mnoha sty lety to tam naši předkové slyšeli. A tak se naše děti biflují, místo toho, aby dělaly něco jiného. Třeba si procvičovaly porozumění textu nebo jak napsat kus textu tak, aby mu bylo rozumět. Pořád historizujeme češtinu.

Myslíte, že se to v budoucnu změní?

Nejsem zastánce žádných revolučních změn, i když jsou ve jménu modernizace, jako „pojďme se vykašlat na vyjmenovaná slova a psát všude měkké, nebo jen tvrdé i“, to je stejná intervence jako u těch obrozenců.

Ale pokud lingvisti začnou popisovat jazyk opravdu takový, jaký je, a ne takový, jak si představují, že by měl být, a pedagogové s tím budou seznamovat děti a studenty tak jako třeba v hodině přírodovědy, kde se děti neučí poznávat jen „krásné“ květiny, ale i kytky ošklivé nebo jedovaté, pak může přijít změna.

Bylo fajn zjistit, že jsem se nezbláznil

Vidíte, my jsme nějak utekli od té vaší osobní historie, jak se z učitele jazyka stal jazykovědec. Učil jste vůbec?

Učil. Ale jak jsem pátral po důvodech, proč mám studenty nutit biflovat se tyhle věci, tak jsem ztrácel motivaci. Nenacházel jsem za tím žádnou vědu, jen emocionální nebo historická

rozhodnutí. Přišlo mi to postavený na hlavu, ale na druhou stranu jsem zjišťoval, že v tom názoru nejsem sám.

Třeba už můj učitel, profesor Petr Sgall, který od 50. let bojuje za zrovnoprávnění obecné a spisovné češtiny, nebo profesor František Čermák, zakladatel korpusu, který vždycky mluvil o tom, že česká lingvistika je příliš preskriptivní a málo deskriptivní. Bylo fajn zjistit, že jsem se nezbláznil.

Kde a jak dlouho jste učil?

Jen rok, na gymnáziu v Troji. Byl to krušný rok, ale jsem za něj rád. Pedagogicky mi to dalo víc než deset let učení seminářů na vysokých školách, jenže přišla nabídka pokračovat ve vědecké kariéře a mě věda táhla, tak jsem školu opustil. Ale stihl jsem si některé věci vyzkoušet v praxi.

Co třeba?

Třeba jestli zvládneme látku bez používání termínů spisovný/nespisovný. V září jsem věnoval dost času popisu různých komunikačních situací – doma, v novinách, v televizním pořadu, v učebnici. Udělali jsme si škálu a pak jsme zařazovali různé jazykové prostředky do různých situací. Dohadovali jsme se o tom, kdy ještě můžeme použít *bysme,* a kdy už to není vhodné. Ukázalo se to jako dobrá cesta.

Při opravování slohů jsem formální chyby označil, ale nepočítal do výsledku. „Tu dvojku máš za to, že se ztratil příběh.“ Jde to, ale musí se chtít, jednodušší je vynucovat jednoznačná formální pravidla, třeba ta vyjmenovaná slova, která mi leží v žaludku. Osvícených pedagogů je pořád málo, ale nejsem didaktik, nebudu říkat učitelům, jak mají učit. Ta změna musí přijít v první řadě od nás, od lingvistů.

Jak se na tom sám podílíte?

Máme spoustu konferencí, seminářů, šíříme povědomí o korpusu a práci s ním. Ale ten trend je započatý. Vydali jsme s kolegy v roce 2010 *Mluvnici současné češtiny* (sundává z poličky knihu), je to první česká mluvnice udělaná podle korpusu. Vidíte tady třeba tvary, které byste v jiné mluvnici nenašla. Máme tu *trpím,* ale i *trpim, prosí* i *prosej, lidmi* i *lidma*.

Etalon korpusových mluvnic je Longman *Grammar of Spoken and Written English* od *Douglase Bibera*, což je ikona korpusové lingvistiky (vytahuje další knihu). Vidíte, všechno má rozdělené na jednotlivé komunikační rejstříky: konverzace, beletrie, žurnalistika a akademická literatura. Pro každý tvar pak udává četnost jeho užití v tom kterém prostředí. Vyšlo to na sklonku tisíciletí, jsme tedy trošku pozadu, ale o moc ne.

VÁCLAV CVRČEK, jazykovědec, zástupce ředitele Ústavu Českého národního korpusu Filosofické fakulty UK, jejímž je absolventem (vystudoval obory čj a literatura, lingvistika–fonetika a postgraduální studium korpusová lingvistika). Pochází z Liberce, žije s rodinou v Praze. Foto: Gabriel Kuchta, Deník N

Co by se studenti měli o jazyku dozvídat především?

Měli bychom jim ukazovat, jak jazyk skutečně vypadá, pak se v něm zorientují a budou ho adekvátně používat. Tento přístup otvírá paletu možností, nezavírá žádné dveře.

I my dva spolu stále dojednáváme formu komunikace, přeskakujeme mezi obecnou a spisovnou češtinou, sám si kalibruju, do jaké míry má být náš rozhovor formální, či neformální, jak chci působit, ale fakticky využívám celé spektrum. A vy zjevně taky. Je to větší bohatství, než kdybychom si řekli: budeme hovořit výhradně spisovnou češtinou, což je zcela běžný předpoklad některých jazykových příruček.

Zanecháváme po sobě jazykové „otisky prstů“

Působilo by to dost nevěrohodně. Ale musíme zde asi říct, že naše povídání vypadá teď jinak než to, které čtenáři čtou. Psaný rozhovor už není autentickým přepisem, protože to, co nám v hovoru nijak nevadí, by komplikovalo srozumitelnost při čtení, když přeskakujeme od tématu k tématu a zase zpátky, nedokončujeme věty, ulítne nám shoda podmětu s přísudkem… To mi připomíná, že český jazykový korpus využívá nahrávky autentických rozhovorů mezi lidmi. Jak je získáváte?

Spolupracujeme s univerzitami v regionech, kde pro nás studenti nahrávají – v rodinách, na kolejích, v hospodách. A pak se s tím materiálem taky učí pracovat a zpracovávat ho. Musíme mít souhlas mluvčích, ale ten často získáme až zpětně, aby autenticita nebyla narušená.

Už samotná nahrávka je archivovaná anonymizovaně, jsou z ní vypípána vlastní jména a věci, které by mohly vést k odhalení identity mluvčích. Ale i když to mluvčí předem vědí, tak se hlídají jen chvíli, dlouho to nikdo nevydrží. Na tom koneckonců staví i forenzní lingvistika, že i když píšete, zanecháváte specifické jazykové stopy.

Musí být docela sranda vidět přepisy těch autentických konverzací, ne? Dovedu si představit ty věty, které nikam nevedou…

Ano. Většina našich popisných konceptů neplatí, vymezení slova nebo věty, bez nichž si jazykový popis neumíme představit, je v případě mluveného jazyka velmi problematické až nemožné. Často mluvíme jen v nějakých úsecích, více či méně propojovaných, bez jasné struktury, nemívá to podmět a často ani přísudek, většinou několikero falešných začátků a vyšinutí z vazby, ale zachraňuje to obrovská redundance.

Ale rozumíme si. Přitom jeden z hlavních argumentů těch, kdo lpí na kodifikaci a normativech pro český jazyk, je, že kdyby si každý psal a mluvil, jak by chtěl, nerozuměli bychom si.

A je to absurdní. Nikdy v minulosti kodifikace nebyla, vzhledem k tomu, jak starý je lidský jazyk, je to relativně novodobá záležitost. Nikdo přitom nedoložil, že by si lidi někde nějak fatálně nerozuměli, pokud mluví stejným jazykem. To je hloupost.

Když mluvíte nesrozumitelně, tak dostanete penalizaci – prostě se nedomluvíte, tak si příště rozmyslíte líp, jak mluvit, protože primárním cílem lidí je dorozumět se. A nějaké snahy à la *když budete poslouchat naše doporučení, domluvíte se lépe* – to je blbost. Žádný lingvista není schopen nahradit to, na co jsme přišli v milionech interakcí. Ale ty hrany se postupně obrušují, věřím, že dochází k erozi toho skalního přístupu k jazyku.

A strhlo se jazykové běsnění

Co může vést ke změně?

Paradoxně v tom může hrát roli onen vynucovací prostředek v jazykové regulaci, *Pravidla českého pravopisu*. Velmi problematická jazyková příručka, kočkopes mezi slovníkem a gramatikou, ani jedno to přitom není pořádně.

Jak to? A jakou roli v tom může sehrát?

První Pravidla vyšla v roce 1902, větší revize proběhla v roce 1957. Pak docházelo k dílčím úpravám až do roku 1993, kdy vyšla Pravidla nová. Ty změny se chystaly od konce 80. let, mnozí lingvisté tehdy radili počkat, protože společnost i jazyk byly v pohybu – ale vydalo se to.

A strhlo se pravopisné běsnění, které nikdo z lingvistů nečekal; vyjádřil se snad každý, lidi s pocity znovunabyté svobody odmítali diktát, jak psát *citron* nebo *filosofie*. Až ministr školství Ivan Pilip svolal schůzi a z ní vzešel tzv. Dodatek, který všechno, co nová Pravidla udělala nově, ponechal v platnosti, ale zároveň stanovil, že platí i to, co předtím.

Pane jo, to jsem už zapomněla, to je vtipný.

A od té doby si nikdo nedovolil vydat nová Pravidla. Je to stále problematičtější a je možné, že už vydaná nebudou. Stávající předpis je statický, jazyk se vyvíjí, ubíhá, v jednu chvíli už budou Pravidla nesnesitelně zastarávající, přestanou být použitelná. Pokud nikdo nebude mít

odvahu tuhle bláznivou příručku nově vydat, je šance, že se to prolomí a psaný jazyk se třeba uvolní.

Jazykové brusičství je náš intelektuální export

Jak je na tom vlastně čeština v té zakonzervovanosti a normativním přístupu ve srovnání s jinými jazyky?

Záleží na síle purismu. Obecně se dá říct, že čím je jazyk méně sebevědomý, tím úzkostlivěji si ho národ střeží, čím sebevědomější jazyk, tím menší potřeba ho regulovat. Britové jako sebevědomá koloniální mocnost k žádné výrazné regulaci nikdy nepřistoupili.

Angličtina je velký jazyk i proto, že volně přejímá z jiných jazyků, dává jí to obrovskou sílu a internacionálnost. Na rozdíl od francouzštiny, která vždycky trpěla asi nějakým pocitem méněcennosti, měla potřebu se vymezit, aby se nerozplynula z německého či anglického vlivu. Ale pořád je to větší jazyk než čeština. Čím menší jazyk, tím ohroženější si obvykle připadá.

A tak máme tendenci ho chránit, uzavírat, konzervovat, stavět kolem něj zdi…

… a být purističtí. Ve slovanském areálu každopádně je to věc poměrně častá. Brusičství jazyka, to je mimochodem náš intelektuální export.

Předpokládám, že nemáte v oblibě tzv. grammar nazis, tedy lidi, kteří cupují každé gramatické „provinění“ na Facebooku… Člověk se snaží vyjádřit, sekne se v *mně/mě* a už je jedno, co chtěl říct, protože diskutéři se zabývají jeho formální chybou.

Opravování formálních chyb patří k typickým argumentům ad hominem. Prostě podpásovka. Rozdělení společnosti je až absurdní. Tohle je bohužel často používaná strategie elit vůči „neelitám“ a to mě mrzí, protože to je pak argumentem pro ty, kdo říkají, že se elity povyšují a jsou arogantní.

Ale je tak svůdné říct: Čím víc hrubek, tím víc „vlastenec“!

Souhlasím. Ale měli bychom mít dostatek věcných argumentů, když s „vlastenci“ nesouhlasíme. Obecně mají lidi o svém vlastním písemném nebo mluveném projevu lepší představu, než je skutečnost. A rádi pak soudí z projevu jiných, jak „jazyk upadá“, to je konstanta. V nějaké mytické době… Tehdy byl jazyk asi ideální.

Přiznávám, že mě občas rozesmějí něčí tendence být naopak tak „spisovný“, až přestřelí a je z toho třeba *cinkání klíčemi.* Nebo se z *už* najednou stane *již,* z *kterého jenž* a z *píšu a miluju* je *píši* a *miluji*.

To se děje často. A podívejme se do korpusu (otevírá stránky www.korpus.cz). V psaném jazyce 20 procent *již* vedle 80 procent *už*, nejvíc pak v akademických textech. V mluveném jazyce 99,9 % pro *už*. Slovo *již* je dnes umělá věc, člověk to neřekne, jak je rok dlouhej, ale hodně lidí má tendenci to použít, když má napsat nějaký trošku formálnější text.

Dobré je vybírat adekvátní prostředek pro danou situaci a publikum, pak se totiž budou čtenáři/posluchači soustředit na to, co jim chcete sdělit, nebude je rušit forma. A rušivé je, když to podceníte, ale i přestřelíte.

Jako s oblečením. Je blbý přijít v tričku a rozervaných džínách mezi lidi v oblecích a šatech, stejně jako přijít ve fraku tam, kde jsou všichni v saku nebo v koktejlkách.

Pokud ovšem k formě nechcete naopak přitáhnout pozornost. Básník záměrně používá nečekané formální prostředky, když budete psát nějaký názor na Facebook, budete spíš chtít připoutat pozornost k obsahu.

Korpus je jazykové safari

Několikrát jsme už zmínili korpus, tedy Český národní korpus, což je elektronická databáze miliard slov z tisíců různých psaných i mluvených textů, formálních i neformálních, z beletrie, odborných článků, publicistiky, ale i z běžné mluvy v neformálních situacích. V čem je korpus jiný, než jsou slovníky?

Jeden z našich sloganů by mohl být *Korpus je jazykové safari*. Můžete se s jeho pomocí vydat dovnitř jazyka a koukat, jak opravdu žije, jak se používá. Každé slovíčko vidíte v jeho autentickém prostředí, kdo ho používá a v jakých situacích, v jakých typech textu. Není to slovník nebo mluvnice, ale lingvisté ho využívají k jejich tvorbě. Jazykový korpus ocení taky nerodilý mluvčí, sám třeba využívám korpus anglický, když hledám vhodný výraz, spojení. Třeba… vzbudit dojem. Dojem je impression…

Projekt dnes sdružuje mnoho různých korpusů určených pro různé účely. Největší z nich se jmenuje SYN a v současnosti je k dispozici ve verzi 7. Ta obsahuje 4,26 miliardy slov psaného jazyka (beletrie, odborné literatury, ale zejména publicistiky, která má vždy největší objem, je nejsnáz dostupná). To odpovídá 325 milionům vět.
Ale *vzbudit – wake up* tam zrovna nepasuje…

Tak se podívám do korpusu, který mi nabídne *to give the impression.* Díky korpusu můžeme poprvé jevy v jazyce kvantifikovat, měřit, což je běžné v přírodních vědách, ale v humanitních tolik ne. Už se máme o co opřít. Je taky vidět, že intuice selhává…

Když se zeptám, jak byste seřadila podle frekvence slova *dívka, děvče* a *holka* v psaném jazyce, můžete pouze hádat, takovouhle věc pomocí intuice odhalit nelze (*pořadí je – dívka 156 výskytů na milion slov, holka 66 výskytů na milion slov a děvče 43 výskytů na milion slov, pozn. red.*).

Vůbec jsem netušila, že u zrodu prvního jazykového korpusu stál Čech, český emigrant z roku 1948 v Americe, Jindřich Kučera. Působil na univerzitě, kde jste byl jako hostující vědec, setkal jste se s ním?

Ano, Henry Kučera, který působil na Brownově univerzitě (stát Rhode Island v USA). Bohužel nesetkal, byl jsem tam až pár let po jeho smrti. Ale je tam jeho knihovna, do konce svého života sledoval dění v Čechách, byla to vědecká celebrita, česká lingvistika jich mnoho nemá. Na Brownově univerzitě je dodnes malý fond Henryho Kučery na podporu lidí, kteří tam přijedou jednorázově přednášet o češtině, chtěl, aby tam česká stopa zůstala i po jeho smrti.

Vytvořil první korpus *Brown Corpus of Standard American English*, dodnes se mu říká Brown corpus, výběr textů v soudobé americké angličtině z roku 1961 z celkem 1000 vzorků. A začátkem 80. let napsal pro IBM taky první spell checker, program pro kontrolu pravopisu, které se pak využívaly v textových procesorech, jako je Word.

„Před pěti lety se nám konečně povedlo vypracovat ze sklepa fakulty až sem.“ Ústav Českého národního korpusu FF UK byl založený v roce 1994, nyní sídlí v Panské ulici, shodou okolností přesně v místech, kde bývala dlouhá léta redakce časopisu Týden (kde pracovala i autorka tohoto rozhovoru). „Konečně máme, řekněme, lidský podmínky k práci, i když je nás třeba v tomhle kanclu pět. Ale máme okno, denní světlo! Krása. I když to pořád není standard západního vědeckého pracoviště,“ říká Václav Cvrček. Foto: Gabriel Kuchta, Deník N

Komu čemu se ubližuje? Rusku

Čím jste se v poslední době vědecky zabýval vy sám?

S kolegyní Fidlerovou právě z Brownovy univerzity jsme dělali analýzu české mutace portálu Sputnik News, což je jeden z celé sítě sputnikovských dezinformačního webů, které jsou oficiálně placené Kremlem, je to přiznané.

Analýzu jsme dělali na datech z vrcholící ukrajinské krize a po ní. Zabývali jsme, jakým způsobem je utvářen obraz Ruska v těchto médiích. Baví mě, když člověk nachází zajímavý detail tam, kde by to nečekal. Zdánlivě nudná věc – frekvence třetího pádu, dativu…

Tedy komu? čemu?

Není to v textech příliš frekventovaný pád, ale v těchto textech bylo právě Rusko neobvykle často prezentováno ve třetím pádu, tedy „Rusku“. Šli jsme do toho hlouběji a zjistili jsme, že skoro všechny výskyty jsou ve smyslu „proti Rusku“ nebo „vůči Rusku“. Vždycky jde o sankce či hrozbu.

Mnohem víc než v mainstreamových médiích se zde objevuje Rusko jako oběť, Rusku se pořád děje něco zlého, někdo mu ubližuje, někdo mu něčím hrozí. Jde o konflikt, kterému se podle Sputniku musí Rusko logicky bránit. A nejde to samozřejmě jinak než vojensky. To už známe, že? A tohle zjistíte tak, že nudně počítáte slova a frekvence. A zjistíte odchylku od běžného úzu a ptáte se proč. Jejich propaganda tedy nefunguje na bázi vychvalování Ruska, ale říkají, podívejte, jak se Rusku škodí.

Co dalšího se ukázalo?

Třeba obraz prezidentů. Zatímco Putin se v textech objevuje spíše v aktivní pozici konatele, nositele děje (tzv. agentu), Porošenko naopak v pasivní pozici objektu, se kterým je nějak manipulováno (tzv. patientu). Sputnik Porošenka prezentuje jako slabého vůdce, jeho jméno se málokdy objeví v prvním pádě, zatímco Putin naopak téměř výhradně v silné aktivní pozici.

Může vaše analýza přispět k tomu, že vznikne třeba algoritmus, podle kterého se snadno odhalí dezinformační text? Dílna, ze které text pochází?

Naše metoda není identifikační, spíše exploratorní, tedy spíše hledáme vysvětlení pro jevy, které jsou jinak než v běžném publicistickém textu, zjišťujeme proč, snažíme se nabídnout interpretaci. Různé propagandistické dílny používají různé prostředky, i když se v lecčems shodují. Obávám se, že jsme zatím daleko od jednoduchého nástroje, protože propaganda s nástrojem jazykové manipulace pracuje na vysoké úrovni.

Rafinovaně?

A komplikovaně. Ale věřím, že poznání, jak funguje Sputnik nebo podobné weby, nás přivede na cestu, jak s tím nějak účinně bojovat. Doufám.

Nikdo se nechce brodit tím bahnem

Není pro lingvistiku výzva hledat třeba jazykové zákonitosti v postech a komentářích trollů, nejsou to roboti, jsou to lidé, kteří musí mít nějaký zautomatizovaný vycvičený „rukopis“, ne?

Koment musí vyprodukovat hrozně rychle, jsou pod tlakem, nemůžou vymýšlet inovativní věty, určitě. Máte pravdu, ale nevím o tom, že by to někdo dělal. Možná to souvisí s tím, že se nikdo nechce brodit tím bahnem debat.

Skoro se divím, že to třeba nerajcuje nějakého doktoranda matematické lingvistiky a neudělá z toho diplomku.

Bylo by to určitě strašně zajímavý, ale je tam problém. Máte jen ta data – komentáře lidí, který se můžou tvářit různě, nepoznáte, kdo z nich je troll, abyste mohla udělat čistý trollí datový balík a z něj extrahovat jazykový profil.

Potřebovala byste jednoznačný soubor, který na začátku nemáte. Ale jistě by to nějak udělat šlo, věřím, že po sobě nějakou stopu zanechávají, je jen otázka stavu našeho poznání, jestli na ni lze přijít. A já myslím, že jo.

Lingvistická analýza se používá i jako nástroj kriminalistický, srovnáním textů – psaných i mluvených – se dá zjistit nejen, jestli to psal stejný člověk, ale jestli je výpověď autentická, nebo naučená, zmanipulovaná, či lživá. A matematická lingvistika je i nástrojem pro odhalení plagiátorství. Poslední případ, o kterém psal i Deník N, je práce historika Kováře. On sám plagiátorství odmítá, ale důkazy svědčí proti němu, z pozice prorektora odstoupil. Vy jste do případu taky přispěl…

Mě to profesionálně zaujalo, ale vůbec jsem nečekal, že můj facebookový status bude mít takový dopad, chtěl jsem na tom jen demonstrovat použití jedné konkrétní statistické metody.

Když studenti odhalili výraznou shodu poznámek pod čarou v práci Berryho Cowarda a Martina Kováře, tedy že se odkazují na stejná díla se stejnou stránkou, napadlo mě se podívat ještě na jejich pořadí v textu. A když se velká většina vašich poznámek pod čarou

shoduje s poznámkami pod čarou nějakého úplně jiného textu a navíc je máte skoro ve stejném pořadí, to už nevysvětlíte náhodou.

Obdivuju ty studenty, kteří na případ ukázali, jak to zvládají. Musí to pro ně být velmi komplikovaná situace, dostali se pod obrovský tlak. A nedokážu si představit, že by někdo tohle udělal jen pro vlastní zviditelnění, z legrace, to už musí být závažné důvody jít s tím ven.

Jazyková analýza a zrušení trestu smrti

Poprvé v historii soud uznal forenzní lingvistickou analýzu jako důkaz v případu Timothyho Evanse, respektive vraždy jeho ženy, v polovině 20. století. Timothy se nejdřív přiznal k vraždě, pak ale změnil výpověď a za vraha označil souseda Johna Christieho. Policie ale uvěří Christieho verzi a popraví Evanse. Pár let po jeho popravě ale vyjde najevo, že John Christie je sériový vrah, je usvědčen z vraždy šesti žen a pak se dozná i k vraždě Evansovy ženy.

Veřejné mínění si žádalo rozluštění. A vtom přichází anglista Jan Svartvik se svou analýzou, která byla z dnešního pohledu poměrně naivní, ale funkční. Měřil tehdy délku vět v jednotlivých výsleších Evanse, vycházel z předpokladu, že čím větší intelekt a zkušenost s formulováním myšlenek, tím delší věty jste schopná tvořit, a naopak.

Čím menší intelekt a sečtělost či rétorická zkušenost, tím kratší věty přirozeně používáte. Evans byl údajně pologramotný prostý muž. Nebylo nakonec s podivem, že části výslechu, kde Evans vypovídal o sobě, o tom, co dělá, kde bydlí, jak žili s manželkou, jsou v krátkých jednoduchých větách, stejně jako jeho změněná výpověď o tom, že vraždu nespáchal on, ale soused Christie, zatímco první doznání bylo formulované ve větách relativně složitých, jakoby naučených, a bylo zjevně zmanipulované.

Chudák Evans, tomu už to nepomohlo.

Bohužel ne, ale byl alespoň rehabilitován. A ten případ je významný proto, že soudní dvůr přijal posudek lingvisty jako důkaz. A jelikož to byla justiční vražda par excellence, začalo se v Británii volat po zrušení trestu smrti. (*K tomu skutečně došlo o pět let poté v roce 1966, pozn. red.*)

Jak se pak vyvíjela lingvistická forenzní analýza?

Od začátku šlo o to, jak najít ten „otisk prstu“ v textu. Ať chcete, nebo ne, vždycky ho po sobě zanecháte. Nejdřív se měřila délka vět a četnost nějaký slov, dnes už se jde na úroveň znaků, třeba interpunkčních znamínek. Máme relativně omezený repertoár interpunkčních znamínek, tečka, čárka, středník, dvojtečka, závorky, uvozovky, a to, jakým způsobem je používáme a jak často, nás do značné míry jako charakterizuje. To by člověk možná ani nečekal.

Vy forenzní analýzu děláte?

Byl jsem párkrát osloven, ale odmítám to. Je to příliš velká zodpovědnost, pořád je to jen statistika. Ve výsledku dostanete nějakou pravděpodobnost, že tenhle člověk je, nebo není autorem toho textu. A já si netroufám jít s nějakou pravděpodobností k soudu a ukázat na „pachatele“. V téhle pozici být nechci. Ale teoreticky mě to moc baví.