

**Posudek oponenta disertační práce Ondřeje Dufka**  
**„Kritická analýza jazykových ideologií v českém veřejném diskurzu“**  
**předkládané v roce 2018 na ÚČJTK FF UK**

**I. Stručná charakteristika práce**

Ondřej Dufek ve své disertační práci využívá postupy korpusově založené analýzy diskurzu pro výzkum jazykových ideologií v českém mediálním a politickém prostoru. Cíle práce jsou explicitně definovány v části 3.1. Vedle popisu ideologií, které jsou používány v diskurzu o češtině, si autor vytyčil také obecnější sekundární cíle, jimiž je zpřesnění vymezení jazykových ideologií a „kultivaci“ metod korpusové lingvistiky využívaných v rámci analýzy diskurzu. Práce má logické rozvržení a postupuje od obecnějších a metodologických úvah o ideologiích, (kritické) analýze diskurzu, korpusových metodách až k samotnému popisu shromážděného materiálu a jeho analýze.

**II. Stručné celkové zhodnocení práce**

Předložená disertační práce prokazuje schopnost Ondřeje Dufka pracovat s odbornými zdroji, kriticky hodnotit teoretické koncepty a metody, aplikovat je na nashromážděný materiál a výsledky analýzy interpretovat. Potud lze konstatovat, že práce splňuje předpoklady, které jsou kladeny na disertační práci. Zároveň ovšem k textu lze mít poměrně velké množství různých výhrad, které jednoznačně pozitivní hodnocení znemožňují a které by mělo být předmětem diskuse v rámci obhajoby.

**III. Podrobné zhodnocení práce a jejích jednotlivých aspektů**

Za pozitiva práce lze označit její přehlednou strukturaci a práci s prameny. Obojí se ukazuje už v úvodní kapitole, kde se představuje rozvržení práce, a dále v kapitole 2 a 3, které sumarizují různé přístupy k jazykovým ideologiím a stanovují cíle výzkumu. V této souvislosti lze snad pouze namítnout, že první kapitola není úvodem do problematiky, tedy zasazením do širších lingvistických či společenských kontextů, ale pouze komentovaným obsahem.

Věcné připomínky mám až ke kapitolám následujícím a v jejich výčtu budu postupovat „chronologicky“, tj. ve shodě se strukturou práce.

O. Dufek oprávněně uvozuje oddíl o klíčivosti (4.2) upozorněním, že proces identifikace klíčových slov předpokládá netriviální metodologická rozhodnutí, která mohou významně ovlivnit analýzu – rozdílné seznamy klíčových slov jistě mohou vést k rozdílným analýzám. Vzhledem k tomu, jak subjektivní proces interpretace textu ze své podstaty musí být, by však bylo na místě posuzovat míru metodologických rizik v relaci k možným úskalím dalších kroků analýzy. V této souvislosti si dovoluji vzpomenout přednášku P. Bakera *Keywords: Signposts to objectivity* (pronesenou na konferenci CADS 2012 v Bologni a částečně i na CL 2014 v Praze), kde kriticky hodnotil svůj výběr z klíčových slov v již publikovaných analýzách, který sloužil jako základ pro interpretaci a který při zpětném pohledu na data autor musel v mnoha ohledech korigovat. Nelze v této souvislosti podle mého názoru opomenout, že míra rigoróznosti a subjektivity je v těchto dvou krocích (identifikaci klíčových slov a následném výběru podmnožiny z nich pro interpretaci) nesrovnatelná a že v druhém z nich je ukryt řádově větší potenciál pro svévolnou či podvědomou manipulaci s interpretací ze strany badatele. Aniž bych tedy chtěl snižovat relevanci snesených námitek proti metrikám měření klíčivosti a metodologii vytváření

seznamů klíčových slov, soudím, že pozornost věnovaná tomuto tématu (zde i v jiných pracích) je neadekvátní tomu, jaká rizika číhají v konsekventních krocích.

V upomínce na výše zmíněnou přednášku P. Bakera nelze rovněž nevzpomenout na jeho upozornění, jak podstatnou roli hrají při analýze neprávem opomíjená funkční slova a nejjobecnější lexikální jednotky. V tomto duchu lze mít celou řadu připomínek i k posuzované práci, která se téměř výhradně zabývá autosémantiky (explicitně viz s. 55).

V opozici k tvrzení O. Dufka na s. 36 se domnívám, že „široká užívanost“ log-likelihood nepramení z toho, že je tato míra implementována v korpusových nástrojích (to je většinou důsledek poptávky ze strany uživatelů); za podstatnější považuju fakt, že log-likelihood je svým zaměřením podobný  $\chi^2$  testu, na rozdíl od něj je však upraven pro užití i s relativně malými hodnotami v kontingenční tabulce.

Drobná faktografická korekce poznámky 31: tematická koncentrace byla v *KWords* implementována od začátku, navzdory tomu, jak se postupně a se zpožděním vyvíjela dokumentace k tomuto nástroji v rámci Wiki příručky ČNK.

Kritika metriky založené na p-value (ať už se jedná o jakýkoli statistický test) na s. 38 je jistě oprávněná, obávám se ovšem, že zde došlo k záměně cíle analýzy (identifikace prominentních slov vypovídajících o obsahu textu) s prostředky k jejímu dosažení. Ačkoli tedy lze souhlasit s tím, že podstatné pro klíčová slova jsou jejich vlastnosti (ve zkratce: prominence), nelze tuto jejich vlastnost operacionalizovat jinak než pomocí frekvence a statistické signifikance rozdílu vyjádřené pomocí p-value. Jinak se totiž může stát, že budeme vybírat jako klíčová také ta slova, jejichž výskyt je ve zkoumaných textech oproti referenčnímu korpusu náhodný.

Jako zásadní metodologickou nevýhodu etablovaného určování keyness uvádí O. Dufek nedostatek pozornosti věnované disperzi. Tuto kritiku, která nezaznívá pouze od něj (viz např. dosud nepublikovaný návrh Egberta a Bibera popsany v letos vydané studii Egber, J. & Schnur, E. 2018: *The role of the text in corpus and discourse analysis*. In: Taylor, C. & Marchi A. (eds): *Corpus Approaches to Discourse. A Critical Review*. New York: Routledge, p. 158–170) považuji za poměrně jednostrannou. Při řešení výzkumného úkolu, který je podobný tomu, jaký si předsevzal O. Dufek, je jistě dobré založit analýzu na těch klíčových slovech, která se vyskytují pravidelně v celém korpusu, resp. diskurzu. V jiných výzkumných úkolech, které rovněž využívají analýzu klíčových slov, to ovšem tak být nemusí. Uvedu příklad z vlastní zkušenosti: literární analýza Tolkienova *Hobita* může významně těžit z faktu, že mezi klíčová slova se propracují i lexémy, které souvisejí s postavou Gluma, který se ovšem v románu vyskytuje významně pouze v jedné kapitole. Analýza, která by rovnoměrnou distribucí klíčových slov striktně vyžadovala, by taková klíčová slova vůbec nepřipustila. Shrnutí: jakkoli je rovnoměrná disperze slova povětšinou důležitým vodítkem, v některých typech analýz je cenná i informace o nevyváženém a přesto prominentním užití určitých lexémů, což je ovšem požadavek, který jedna univerzální metoda identifikace klíčových slov nemůže uspokojit.

Drobné nepřesnosti se objevují v popisu effect-size metriky DIN na s. 40. Jednak není pravda, že ve vzorci oproti %DIFF není násobení stem, jak stojí v textu, a zadruhé neplatí, že pro slova, která v referenčním korpusu nenajdeme, vychází  $DIN = 0$  (v takovém případě je  $DIN = 100$ , jak se správně uvádí v poznámce 41).

Celkově k metodologii určování klíčových slov: Ačkoli nelze nic namítat proti volbě metriky (kombinace log-likelihood jako testu statistické významnosti a Cohenova  $d$  jako ukazatele velikosti účinku), práce se konec konců neprezentuje jako statistická, ale především jako diskurzivně-analytická, a zvolené míry jsou běžně v pracích tohoto zaměření užívány, přeci jen bych očekával diskuzní poznámku ke Cohenovu  $d$  (na s. 41). Jeho blízký vztah k  $t$ -statistice předpokládá dodržení některých parametrů zkoumaných dat, zejména normalitu jejich rozdělení, což většinou v korpusových výzkumech splněno není (jak na jiném místě konstatuje i autor). Ačkoli je  $t$ -statistika robustní a vůči porušení některých svých předpokladů poměrně imunní, použití Cohenova  $d$  se právě v nerespektování základních předpokladů blíží asociální míře  $t$ -score, kterou lze z těchto důvodů považovat za překonanou. V této souvislosti se proto nabízí otázka, proč namísto Cohenova  $d$  nepoužít např. Cramerovo  $V$  nebo Phi ( $\phi$ ) koeficient, které mají analogický vztah k  $\chi^2$  rozdělení, jako má Cohenovo  $d$  k  $t$ -rozdělení, s tou výhodou, že  $\chi^2$  normalitu dat nepředpokládá a použitý log-likelihood test má  $\chi^2$  rozdělení.

Další nepřesnost se objevuje v popisu  $h$ -bodu na s. 43, kde je uvedeno, že dělí frekvenční křivku na dvě poloviny. Ve skutečnosti je oblast nad  $h$ -bodem (synsémantická) typicky mnohonásobně menší (obsahuje méně typů), než oblast pod ním (autosémantická).

Autor považuje za hlavní výhodu zjišťování tematické koncentrace to, že nevyžaduje referenční korpus. To přitom není pouze technologické zjednodušení analýzy; má to zásadní konsekvence pro interpretaci. Při určování klíčových slov je totiž třeba mít na paměti, že jejich prominence je vždy vázána na srovnávací základ (referenční korpus). Není-li žádný takový referenční rámec určen, nelze v podstatě interpretaci k ničemu vztáhnout (pro koho a v jakém smyslu je dané slovo prominentní). Dovolil bych si dokonce poněkud apodikticky tvrdit, že bez (modelového) čtenáře reprezentovaného referenčním korpusem, nemá smysl z hlediska interpretace prominenci měřit.

Součástí disertačního projektu bylo vytvoření dat reprezentující diskurz o jazyce v prostředí parlamentu a v českých médiích. Celkově považují tuto pasáž (kap. 5) za neproblematickou, mám zde pouze několik dílčích připomínek.

- S ohledem na možné další využití nashromážděných dat lze vyjádřit určitou lítost nad nekompromisním čištěním materiálu od poznámek různého druhu. Nejsem uživatelsky obeznámen s nástrojem #LancsBox, v němž byla analýza prováděna, z jeho uživatelského manuálu lze ovšem vyčíst, že je schopen importu dat v XML. Pokud by autor korpusu zvolil tento formát, bylo by možné všechny poznámky v korpusu zachovat, což by mohlo být cenné pro pozdější uživatele.
- Z popisů na s. 51 jsem nepochopil, jaký je rozdíl mezi „zprávou“ a „článkem“ na jedné straně a „komentářem“ a „článkem“ na straně druhé. Zároveň se ovšem nabízí otázka, proč tuto anotaci autor do korpusu přidával, když s nimi v interpretaci významněji nepracuje.
- Pro morfologickou anotaci a lemmatizaci byl použit nástroj TreeTagger, který je implementován v LancsBoxu; je to sice pochopitelné, ale mnohem lepších výsledků by autor dosáhl s využitím modernějších analyzátorů jako je MorphoDiTa, resp. UDPipe.

Analýza vytvořeného korpusu je prezentována v kapitole 6. K její rámcové koncepci křížové analýzy jednotlivých subkorpusů nemám námitek. Zdá se být dobře promyšlená a pokouší se o nastínění různých pohledů, ačkoli by textu prospělo, kdyby interpretace byla založena na jasné hypotéze, kterou má porovnání s konkrétním referenčním korpusem testovat (v závěru se mluví o „plastičtějším pohledu na data“ (s. 99) bez bližšího vysvětlení). Pro čtenářovu lepší orientaci by bylo rovněž dobré zopakovat, s jakými parametry bude identifikace klíčových slov probíhat (hladina významnosti, hodnoty Cohenova

d, viz s. 45). Postrádám pouze informaci o prahové hodnotě minimální frekvence. Jednotlivé výhrady lze mít rovněž ke konkrétním výsledkům:

- V tabulkách klíčových slov jsou vyznačeny „irelevantní“ jednotky – v některých případech jsou takto označeny i lematizační chyby, jindy nikoli (např. lemma *ústav* v tab. 6.1, *etnických* v tab. 6.3), někdy jsou jako irelevantní označena gramatická slova jindy nikoli (např. lemma *být* v tab. 6.10)
- V některých případech bych čistě intuitivně zaměřil pozornost na jiná klíčová slova, např. lemma *jaksi* (s. 61) či lemma *můj* (s. 82) skýtají možnosti analýzy z hlediska postojů a distance k různým konceptům.
- Analýza je doprovázena hojnými (někdy se opakujícími) citacemi; matoucí je, že některá klíčová slova jsou demonstrována i na promluvách, které nepatří do zkoumaného (sub)korpusu. Domnívám se, že to je případ i cit. 10 na s. 62 a cit. 12 na s. 63 poslankyně Svobodové za KSČM, jejíž promluva je uvedena v pasáži analyzující korpus JazParl\_PROTI, ačkoli patří patrně do skupiny PRO. Zároveň je někdy počet ilustrativních citací až příliš rozsáhlý, to je podle mého soudu i případ na s. 89.

Obecněji k analýze materiálu v kapitole 6 přidávám tyto komentáře a připomínky. Výsledky každého porovnání jsou sumarizovány pomocí tabulek a tematického členění. Nabízí se přitom v práci nezodpovězená otázka, jaký je smysl takové kategorizace, když se s nimi dál v interpretaci významněji nepracuje, a nebo dokonce když konkordanční analýza naznačí, že domnělá sémantická skupina je tvořena pouze lexémy užitými v přenesném významu (viz rozbor skupiny „tělo“ na s. 72).

Ke zvážení přidávám návrh na doplnění kolokační analýzy o analýzu tzv. KWLinks, tedy vztahů mezi klíčovými slovy na delší vzdálenost (nikoli v rámci syntagmatu). Např. analýza adj. *spisovný* vyjeví v rámci kolokací jen poměrně očekávatelná doplnění, zaměření na širší kontext by mohlo pomoci v odhalení hlubších axiologických vztahů spojovaných s tímto jinak problematickým pojmem.

K hlavní zjištění analýzy (6.2) bych rád přidal svoji alternativní interpretaci zjištění, že „[p]oslanci, kteří hlasovali proti návrhům jazykových zákonů, se s podporovateli těchto návrhů navzdory protikladnému chování shodují víc, než by odpovídalo (reálným hlasováním podložené) rozdělení na tyto dvě skupiny. Mnoha z nich byly společné jazykové ideologie jako např. jazykový axiologismus (jazyk má výjimečnou hodnotu), protektivismus (jazyk potřebuje neustálou ochranu) nebo purismus (domácí prostředky jsou lepší než cizí)“ (s. 97). Není možné tuto míru shody vysvětlit prostě jenom argumentační strategií? Namísto kategorického nesouhlasu nabízejí odpůrci návrhu (pragmaticky) částečnou shodu ve východiscích a svůj nesouhlas prezentují pouze jako odmítnutí navrhovaných legislativních nástrojů.

Po formální stránce hodnotím práci jako uspokojivou. Nedostatky, které jsem v práci objevil, uvádím pro případ, že by text byl upravován pro další publikování:

- V celé kapitole 2 (Jazykové ideologie) se patrně vlivem dodatečného přečíslování opakuje odkaz na oddíl Klíčové vlastnosti jazykových ideologií s neexistujícím číselným označením (2.1.4. namísto 2.4). Podobně se odkazuje na oddíl 2.2 odkazem na 2.1.2.
- na s. 41, 1. odst. v odd. 4.2.6: přecházejících > předcházejících
- s. 46, 3. odst.: včetně ideologií s těmito skupinami spojených > včetně ideologií s těmito skupinami spojenými
- na s. 63, v 3. odst. je nedokončená věta „..., jež se v konkordancích.“
- s. 71, 4. odst.: závěru > záměru

- s. 79, cit. 40: není jasné, k čemu se autorská oprava „[OD]“ má vztahovat
- s. 86, poslední odst.: přísnější kritérium log-likelihood by značilo použití hodnoty 6,63 namísto 3,84 a nikoli naopak, jak je v textu uvedeno.
- s. 98, 4. odst.: vytkhl > vytkl
- s. 100, 3. odst.: za jednou větou chybí tečka, za jinou jedna přebývá

#### IV. Dotazy k obhajobě

Vedle vyjasnění problematických bodů uvedených výše bych rád autora práce vyzval k explicitnímu zodpovězení těchto dotazů během obhajoby.

Není mi jasná poznámka na s. 39 v bodu 4, kde se probírá možnost využití Mann-Whitneyho testu pro zjišťování klíčivosti, který by zohledňoval disperzi v jednotlivých textech. Na konci odstavce se uvádí, že test se „obejde bez arbitrární hodnoty dostatečnosti“.

Za značně matoucí považuju systematické uvádění referenčního korpusu při analýze kolokací. V obr. 6.2 a 6.3, stejně jako v tab. 6.4 a 6.5 na s. 65 a ve všech dalších analogických tabulkách se uvádí referenční korpus. Autor by měl tento aspekt vysvětlit, protože role referenčního korpusu v analýze kolokací mi není bez dalšího vysvětlení jasná.

K výsledkům analýzy: Interpretace dospívá k poměrně očekávanému zjištění, že diskurz o češtině do značné míry ovládá protekcionistický přístup. Tento postoj přitom není v české jazykovědě neznámý a podle některých autorů do značné míry formuje celou českou kodifikační praxi. Narážím zde zejména na koncept „Syndromu národního údělu“ (Starý, Z. 1995. *Ve jménu funkce a intervence*, Praha: Karolinum, s. 64), který podle Starého fungoval jako fantom jazykové regulace od Národního obrození do současnosti a který povstal z negativního hodnocení jazykového vývoje češtiny v baroku (v tomu viz Stichův text ve Sgall, P. et al. 1992: *Variation in language*, J. Benjamins, kap. 3). Lze mezi tímto konceptem Z. Starého a současným diskurzem, jak ho popisuje předkládaná analýza, vysledovat nějaké paralely či dokonce kauzální vztahy? Jinými slovy, do jaké míry se na prominentním postavení protekcionistických postojů k jazyku podepsala jazykově regulační praxe posledních 200 let?

#### V. Závěr

Z výše uvedeného relativně dlouhého výčtu připomínek a komentářů vyplývá, že k práci mám celou řadu různě závažných výhrad. Většina z nich ovšem nemá povahu zásadních metodologických nebo věcných nedostatků, soudím proto, že by práce měla být obhajována. Po porovnání kladů a záporů práce se proto přikláním k následujícímu závěrečnému konstatování:

Předložená disertační práce splňuje požadavky kladené na disertační práci, a proto ji doporučuji k obhajobě a v předběžně ji klasifikuji jako *prospěl*.

8. května 2018

Václav Cvrček