

Jan Chromý

Práce

s empirickými daty

Příručka pro studenty

Bc. studia ČJL

Práce s empirickými daty

Příručka pro studenty Bc. studia ČJL

Jan Chromý

Grafická úprava Jan Šerých

Sazba DTP Nakladatelství Karolinum

Vydání první

© Univerzita Karlova v Praze, 2014

© Jan Chromý, 2014

ISBN 978-80-246-2801-1 (online : pdf)

Příručka vznikla v rámci projektu Rozvoj a inovace bakalářského studia českého jazyka v Ústavu českého jazyka a teorie komunikace FF UK v Praze (CZ.2.17/3.1.00/33275), který byl realizován v období 2011–2013 za finanční podpory Operačního programu Praha Adaptabilita.



Univerzita Karlova v Praze
Nakladatelství Karolinum 2014

<http://www.cupress.cuni.cz>

Obsah

Úvod	7
1. Dotazník v empirickém výzkumu	10
1.1 Rozsah a formát dotazníku	10
1.2 Instrukce v dotazníku	11
1.3 Otázky v dotazníku	13
1.4 Předvýzkum a pilotáž	16
1.5 Administrování dotazníků	18
1.6 Co dělat s vybranými dotazníky?	20
2. Proměnné a jejich vztahy	22
2.1 Typy proměnných	22
2.2 Vztahy mezi proměnnými	25
2.3 Hypotézy	26
3. Koho či co zkoumáme?	28
3.1 Vzorek a populace	28
3.2 Kvótní výběr	29
3.3 Náhodný výběr	30
3.4 Účelový výběr	31
4. Jak zjišťujeme vztahy mezi proměnnými?	33
4.1 Srovnávání statických skupin	34
4.2 Předběžné a následné pozorování jedné skupiny	35
4.3 Klasický experiment	36
4.4 Následné pozorování na dvou skupinách	37
4.5 Reliabilita a validita výzkumu	38
4.6 Longitudinální výzkum	40

5. Redukce a zkreslení	42
5.1 Redukce prováděné ve výzkumu	42
5.2 Typy zkreslení z hlediska vztahu proměnných	44
5.3 Zkreslení z hlediska toho, co nezohledňujeme	46
6. Další pojmy	49
6.1 Indukce a dedukce	49
6.2 Teorie, rámec, model	50
6.3 Výzkumná etika	51
7. Další nástroje sběru dat	54
7.1 Rozhovor jako nástroj sběru dat	54
7.2 Pozorování	56
7.3 Případová studie	59
8. Něco k statistice	60
8.1 Deskriptivní statistika	60
8.2 Pravděpodobnost v inferenční statistice	65
8.3 Rozdělení dat	67
8.4 Jak zjišťovat pravděpodobnost?	72
Literatura	75

Úvod

Tento text slouží primárně jako základní studijní materiál k předmětu *Práce s empirickými daty*, vyučovanému v rámci odborného modulu bakalářského studia oboru Český jazyk a literatura na Filozofické fakultě Univerzity Karlovy v Praze. Příručka pokrývá základní oblasti (především kvantitativně orientovaného) empirického bádání zaměřeného na jazyk a jeho užívání a mohou ji tak využít i další zájemci o tento výzkumný směr.

Základem textu je výklad o jednotlivých pojmech a přístupech empirického výzkumu. Rozsah příručky je relativně omezený a některé otázky tak budou zpracovány pouze obecně. V takových případech však budou čtenáři odkázáni na další, doplňkovou literaturu, z níž se o tématu budou moci dozvědět více.

Východiska

Předkládaná příručka pochopitelně není prvním textem, který se věnuje problematice empirických metod – ať už v jejich obecnosti, anebo v rámci bádání o jazyce. Ve výkladu tak citujeme publikace, z nichž vycházíme, avšak pouze v případě, že se jedná o poznatek, který jiné publikace neobsahují.

Mezi práce, z nichž vycházíme nejvíce a které doporučuji čtenářské pozornosti, spadají následující tituly. Obecnou metodologii empirického výzkumu v sociálních vědách představuje kniha Miroslava Dismana *Jak se vyrábí sociologická znalost* (Disman, 2002). Tato publikace sice jednotlivé problémy ukazuje na sociologickém materiálu – většina poznatků je však přímo aplikovatelná i na výzkum lingvistický, přičemž její velkou výhodou je, že je psána v češtině. Druhou česky psanou příručkou, kterou lze

doporučit, je kniha Jana Volína *Statistické metody ve fonetickém výzkumu* (Volín, 2007). Jak název napovídá, jedná o text zaměřený nikoliv na empirický výzkum jako takový, ale primárně na jeho část analytickou, na kvantitativní vyhodnocení sebraných dat. Knih o statistice je i v češtině řada, právě Volínův text je však vhodné doporučit z toho důvodu, že všechny statistické procedury ukotvuje na fonetickém materiálu, který by měl být studentům lingvistiky srozumitelný.

Z anglicky psaných příruček je třeba vyzdvihnout práci Sebastiana Rasingera *Quantitative Research in Linguistics: An Introduction* (Rasinger, 2008). Tato publikace se v hrubých obrysech orientuje na tutéž problematiku jako práce Miroslava Dismana, její výhodu pro studenty lingvistiky však lze spatřovat v tom, že je založena na jazykovém materiálu. Zároveň je Rasingerova kniha doplněna o návody základních statistických výpočtů v programu Microsoft Excel, což je zvláště užitečné pro ty studenty, kteří nemají přístup k drahým aplikacím pro statistickou analýzu, jako je například SPSS nebo Statistica, anebo se necítí být dostatečně schopní na relativně náročnou práci s programem R (Gries, 2013).

Takto na úvod můžeme zmínit ještě dvě publikace, a to příručku Zoltána Dörniece a Tatsuyi Taguchiové *Questionnaires in Second Language Research: Construction, Administration, and Processing* (Dörniece – Taguchi, 2010) a sborník Lii Litosselitiové *Research Methods in Linguistics* (Litosseliti, 2010). První uvedená publikace se velice podrobně zaměřuje na dotazníkový výzkum v lingvistickém výzkumu, konkrétně ve zkoumání osvojování druhého jazyka. Na závěr knihy je uveden i praktický bodový seznam jednotlivých kroků, jež je třeba v dotazníkovém výzkumu udělat. Uvedený sborník zase základně představuje jednotlivé konkrétní metody empirických lingvistických analýz.

Základní koncepce příručky

Knihy, které jsme uváděli v předchozím oddíle, ale i další texty, které lze o empirických metodách (ať už v lingvistice nebo v jiných sociálních vědách) najít, postupují – logicky – od nejzákladnějších pojmů k těm složitějším. Moje zkušenost z výuky založené na tomto postupu však ukázala problémy spočívající v tom, že „nejzákladnější“ pojem nemusí být pro čtenáře, respektive studenty pojmem nejintuitivnějším. Je tak mnohem pravděpodobnější, že student bude schopen promyslet a pochopit koncepci dotazníku, než například rozdíl mezi nulovou a pracovní hypotézou anebo rozdíl mezi jednotlivými typy proměnných.

Problém navíc tkví v tom, že jednotlivé základní pojmy je zpočátku třeba si prostě zapamatovat bez patřičného ukotvení, přičemž k pochopení toho, k čemu ty pojmy jsou a proč je nutné si je zapamatovat, dochází až o mnoho stránek dál.

V této příručce se tak pokusíme o přístup odlišný. Začneme u nejintuitivněji pochopitelných jevů a postupně budeme odkrývat pojmy a souvislosti další, méně intuitivní, avšak „základnější“. Pevně věřím, že tento přístup bude ku prospěchu věci a studenty od tohoto probíraného tématu neodradí.

Cílem této příručky je v hrubých rysech nastínit základní aspekty empirického výzkumu v lingvistice. Pro úspěšnou realizaci vlastního výzkumu je však třeba zde získané poznatky ještě dále rozvíjet.

1. Dotazník v empirickém výzkumu

Již v první kapitole jsme mnohokrát použili přídavné jméno *empirický*. Ačkoliv to snad není adjektivum zcela neznámé či nepochopitelné, je vhodné si ujasnit, co se tím míní. Nuže, ve výzkumu se vždy snažíme porozumět a popsat svět kolem nás.

Empirický výzkum je pak založen na tom, že poznatky o světě kolem nás získáváme na základě přímého či nepřímého pozorování. Pozorování je zde chápáno velice široce a zahrnuje vše od laboratorních experimentálních výzkumů přes dotazníkové výzkumy a rozhovory až po velmi volné zúčastněné pozorování v určitém společenství. Na základě tohoto pozorování pak získáváme určitá data, která následně kvantitativně či kvalitativně analyzujeme.

Jednou ze základních technik získávání dat (resp. informací) je právě dotazník. Jak všichni víme, princip dotazníku spočívá v tom, že máme určitý předem stanovený formulář (tištěný či elektronický), do něhož se zapisují odpovědi. To, jak dotazník vypadá, má smysl nazírat ze tří pohledů: a) z pohledu rozsahu a formátu; b) z pohledu instrukcí pro respondenty; c) z pohledu znění otázek.

1.1 Rozsah a formát dotazníku

Zásadní poučka při vytváření dotazníku spočívá v tom, že dotazník by neměl být příliš dlouhý. Je-li dlouhý, je velmi pravděpodobné, že se respondent v průběhu vyplňování unaví, začne nudit či naštvě a výsledné odpovědi nebudou podle pravdy, anebo dokonce že respondent dotazník nevyplní celý. Obecně, zdá se, platí, že čím delší dotazník, tím nižší věrohodnost celého výzkumu. Aby však byl výzkum kvalitní a skutečně

něco o světě kolem nás vypovídal, potřebuje badatel co nejvíce odpovědí. Jinými slovy, existuje zde určité pnutí – na jedné straně bychom měli mít otázek co nejméně, abychom předešli nepravdivosti odpovědí, případně tzv. úmrtnosti respondentů (respondenti dotazník nedovyplní), na druhé straně potřebujeme otázek co nejvíce, abychom měli podrobná data, která můžeme analyzovat. Obecně je těžké říci, jaký je ideální rozsah dotazníku – je však zřejmé, že nejdůležitějším měřítkem je čas, který potřebují na vyplnění ti nejpomalejší respondenti. Obvykle se uvádí, že by vyplňování standardního dotazníku nemělo přesáhnout 25–30 minut. Přihlédnout bychom rovněž měli k míře zábavnosti dotazníku – je-li dotazník zábavný, můžeme předpokládat, že respondenti budou odpovídat podle pravdy i po delší době, zatímco v případě, že jsou otázky náročné a respondenti se budou nudit, je vhodné zvolit spíše kratší dotazník. Ve specifických případech se používají i delší dotazníky – v takových případech je však dobré zvážit finanční odměnu za jejich vyplnění.

Vedle délky dotazníku není dobré podceňovat ani to, jak dotazník vypadá, tedy jeho formát. Je velmi důležité, aby dotazník vypadal hezky, respektive profesionálně. Uvádí se, že grafická profesionalita dotazníků vede k vyšší návratnosti a větší poctivosti respondentů při vyplňování. Dotazník by měl být „vzdušný“, tj. mělo by být užito velkorysejší řádkování, měl by mít dostatečně velké okraje, je rovněž dobré graficky odlišit jednotlivé části. K důvěryhodnosti dotazníku rovněž přispívá uvedení loga instituce, v rámci níž děláte výzkum (tj. například loga Filozofické fakulty Univerzity Karlovy v Praze, máte-li svolení k jeho využití).

Celkově platí, že pokud je dotazník z hlediska formátu či rozsahu udělaný špatně, je velká pravděpodobnost, že ve vašich datech dojde k takzvanému zkreslení. To znamená, že výsledky budou ovlivněny něčím, co sami nezkoumáte (např. únava respondenta anebo jeho bezmyšlenkovité zaškrťávání odpovědí, „jak mu přijdou pod ruku“ atd.). Ke zkreslení se ještě vrátíme později (viz kap. 5).

1.2 Instrukce v dotazníku

Klíčovou součástí dotazníku jsou instrukce, které respondenti dostanou pro jeho vyplnění. Rozlišujeme přitom mezi instrukcemi obecnými a konkrétními.

Obecné instrukce musí být uvedeny na začátku každého dotazníku a měly by dotazník podrobně představovat. Z hlediska vědecké etiky je

nutné uvést v dotazníku následující informace: a) kdo konkrétně realizuje daný výzkum (vědecká instituce, soukromá firma, nějaká fyzická osoba?); b) čím se daný výzkum zabývá; c) jakým způsobem se bude pracovat s daty, kdo k nim bude mít přístup; d) míra anonymity.

S těmito informacemi mohou být spojeny určité problémy. Představme si například, že budeme chtít v rámci dotazníkového výzkumu zkoumat přijatelnost určitých slovosledných variant v češtině. V takovém případě je značně kontraproduktivní, pokud bychom respondentům řekli, že zkoumáme právě tyto varianty, protože by to mohlo značně ovlivnit výsledky. Z toho důvodu v takových případech obvykle postačí, když je v obecných instrukcích uvedeno něco zcela obecného – v našem případě například to, že se v rámci dotazníku „zkoumá pohled českých mluvčích na současnou češtinu“.

Druhým problémem je otázka **anonymity výzkumu**. Ne vždy je možné, aby byl výzkum zcela anonymní. V tomto smyslu je vhodné rozlišit anonymitu povrchovou a hloubkovou. Výzkum je povrchově anonymní, když nezachycuje informace, které by vedly k okamžité identifikaci daného respondenta – jinými slovy, když v dotazníku není vyžadováno jméno respondenta, jeho e-mail atd. (povrchová anonymita je pochopitelně narušována mnohem více například u rozhovorů – viz oddíl 8.1). Hloubkově anonymní je výzkum tehdy, když není možné respondenta dohledat ani na základě nepřímých údajů (například si představme, že budeme zkoumat prostřednictvím dotazníků šikanu na středních školách; na určité obchodní akademii bude třída, v níž je 25 děvčat a dva chlapci, z nichž jeden je o rok starší než druhý; pokud dotazník bude mít položky pohlaví a věk, je v případě těchto chlapců snadno dohledatelné, kdo odpovídal jak). Anonymitu respondentů pochopitelně musí zachovávat závěrečná výzkumná zpráva (ať už jde o studii, diplomovou práci či třeba prezentaci), ale je třeba ji uchovávat i v rámci dat, která shromáždíme (tj. respondent musí vědět, komu se jeho dotazník může dostat do rukou – představme si, že by se například dotazníky z výzkumu šikany z uvedené obchodní akademie dostaly učitelům této školy, kteří znají písmo jednotlivých žáků).

Vedle informací, které jsou z hlediska vědecké etiky povinné, bývá v obecných instrukcích vhodné uvést i další informace: a) informace o tom, jak se dotazníky sbírají; b) poděkování; c) kontakt na badatele; d) informace o tom, jak se dostat k výsledkům výzkumu; e) poučení, že v dotazníku nejsou správné a špatné odpovědi a aby respondenti odpo-

vídali výhradně podle svého přesvědčení (toto poučení je užitečné tam, kde můžeme mít oprávněný pocit, že by respondenti mohli odpovídat v souladu s určitou sociální normou apod.).

Kromě obecných instrukcí jsou v dotaznících důležité ještě **konkrétní instrukce**. Ty se vztahují přímo k jednotlivým otázkám a respondentovi říkají, jak je správně vyplnit. Jinými slovy, jedná se o vysvětlení toho, jak odpovídat na určitou konkrétní otázku v dotazníku. Vhodné bývá uvést příklad toho, jak vyplňovat, anebo prostě poučení o tom, zda vybranou odpověď označit křížkem či zakroužkovat, zda lze uvést pouze jednu, anebo více odpovědí atd. Může se to zdát zbytečné, ale je opravdu potřeba uvádět i věci, které považujete za samozřejmé. Až budete realizovat dotazníkový výzkum, sami uvidíte, jak tvůrčím a překvapivým způsobem jsou respondenti schopni k vyplňování dotazníku přistoupit.

Podcenění role instrukcí má obvykle velmi neblahé důsledky pro váš výzkum. V prvé řadě v souvislosti s tím klesá návratnost dotazníků – řada respondentů vám odmítne dotazník vyplnit, protože jim váš výzkum připadá nedůvěryhodný. Zároveň se velice pravděpodobně může stát, že budou dotazníky špatně vyplněny. A pochopitelně – pokud neuvedete informace, které je z hlediska vědecké etiky nutné uvést, vystavujete se nebezpečí, že budete obviněni z jejího porušování, což může vést v nejzazším případě až k vyloučení ze studia.

1.3 Otázky v dotazníku

Tou nejzásadnější částí dotazníku jsou pochopitelně jednotlivé otázky. Jejich sestavení není jednoduché už jenom proto, že je omezený jejich počet, jak jsme již uvedli výše. Zeptat se můžete na cokoli a jakkoliv, ale v reálném dotazníku se můžete ptát pouze na něco a nějak. Existují různé strategie, jak otázky vybrat. Dörniey a Taguchiová (2010) jako nejlepší možnost uvádějí přístup, kdy si sepíšete všechny otázky, které vás k tématu napadnou, ty si poté utřídíte do jednotlivých tematických bloků a pak z každého bloku vyberete několik otázek, které jsou nejpodstatnější a které by měly odkrýt nejpodstatnější informace pro váš výzkumný cíl (respektive pro vaše hypotézy – viz oddíl 2.3). Jiná strategie spočívá v tom, že nejprve s potenciálními respondenty vedete předběžné neformální rozhovory, z nichž vyplyne, jaké otázky by z hlediska vašeho výzkumného cíle mohly být pro danou populaci nejvhodnější, co jsou skutečné problémy, co naopak problémy nejsou atd.

Obecně platí, že v dotaznících musíte u jednotlivých otázek uvést i možnosti odpovědí. Rozlišují se tak otázky otevřené a uzavřené. **Otevřené otázky** spočívají v tom, že je možnost odpovědi volná – respondent zkrátka odpovídá vlastními slovy. Z teoretického hlediska takové otázky poskytují nejlepší, tj. nejpravdivější a nejpřesnější, odpovědi. Prakticky jsou však spojeny s určitými problémy. Především, otevřené otázky mohou respondentům zabírat poměrně dost času, což může vést i k tomu, že respondenti mají tendenci odpovídat na ně relativně zkratkovitě, případně je i úplně přeskaovat. Velký problém je rovněž spojen s tím, že odpovědi na takové otázky se jen velmi obtížně dají zobecnovat, a to zvláště v případě, že dotazník distribuujete velkému množství lidí. Jinými slovy, lze jen velmi těžko objektivně říci, že osoba A a osoba B mají na věc stejný názor, když každá odpoví vlastními slovy. Lze tak říct, že otevřeným otázkám má smysl se v dotaznících spíše vyhýbat, respektive je používat pouze jako doplňkové. Vhodné je jejich použití jako rozvedení odpovědi na předchozí, uzavřenou otázku (např. „Prosím, odůvodněte svou odpověď na předchozí otázku.“).

Uzavřené otázky spočívají v tom, že respondenti vybírají odpověď z nabízených možností. Existuje několik základních typů: a) škály; b) výběr z více možností (tzv. „multiple-choice“ otázky); c) řazení odpovědí; d) odpovědi ano-ne (tzv. „true-false“ otázky); e) číselné položky.

V prvním uvedeném případě respondent odpověď vybírá vyznačením určitého úseku na určité předem dané škále. Rozlišují se především Lickertova a Osgoodova škála. **Lickertova škála** (v klasické podobě) spočívá v tom, že respondenti vyjadřují míru souhlasu s určitým tvrzením. Typicky je pětistupňová (obvykle 1. zcela souhlasím, 2. spíše souhlasím, 3. ani jedno, 4. spíše nesouhlasím, 5. zcela nesouhlasím), někdy však může mít smysl ji formulovat čtyřstupňově, tj. bez prostřední, neutrální možnosti (pak se jedná o takzvanou „vynucenou volbu“). **Osgoodova škála** (jinak též **sémantický diferenciál**) je vymezena dvěma protiklady, přičemž většinou jde o přídavná jména (např. Jak hodnotíte náročnost uvedeného textu? *obtížný ... lehký*), někdy o substantiva (např. Jak na vás z hlediska svého projevu působil tento mluvčí? *odborník ... laik*), ale i příslovce či slovesa. Je důležité, abychom si u tohoto typu škál dávali pozor na použití prostého záporu – mohlo by se totiž stát, že by výsledná škála byla vychýlená (srov. například škálu *lehký ... nelehký*, která je vychýlená, protože adjektivum *nelehký* není na rozdíl od *obtížný* či *těžký* přímý protiklad slova *lehký*).