

Vysoká škola ekonomická v Praze

Bakalářská práce

Vysoká škola ekonomická v Praze
Fakulta podnikohospodářská
Studijní obor: Podniková ekonomika a management



Název bakalářské práce:

Genderové stereotypy o povolání v přirozeném jazyce

Autor bakalářské práce: Šárka Jadviščoková

Vedoucí bakalářské práce: Ing. Mgr. et Mgr. Štěpán Bahník, Ph.D.

P r o h l á š e n í

Prohlašuji, že jsem bakalářskou práci na téma
„*Genderové stereotypy o povolání v přirozeném jazyce*“
vypracovala samostatně s využitím literatury a informací,
na něž odkazuji.

V Praze dne: 12. května 2021

Podpis

Název bakalářské práce:

Genderové stereotypy o povolání v přirozeném jazyce

Abstrakt:

Bakalářská práce je zpracována jakožto replikační studie článku o genderových stereotypech o povolání v přirozeném jazyce. Hlavními oblastmi zkoumání (tj. korpusy) jsou mluva mířená dětem, mluva dospělých a dětí a knihy a audiovizuální média pro děti i dospělé. Cílem práce je detekování asociací mezi jednotlivými profesemi a genderem implicitně zakotvených v textu. V teoretické části jsou popsány pojmy týkající se big dat a machine learning, replikace a replikační krize, genderu a stereotypů a použité analýzy. V praktické části jsou za použití word embeddingů (konkrétně modelu fastText skipgram) analyzovány dané korpusy a výsledné word embeddingy jsou interpretovány pomocí metod meta-analýzy v statistickém softwaru R. Výsledkem práce je nalezení empirického důkazu o genderových stereotypech o povolání v přirozeném jazyce a jejich korelace se skutečnými administrativními statistikami o zastoupení daného genderu v jednotlivých profesích. Práce potvrzuje, že existují konzistentní rozdíly ve vnímání genderu a povolání a jejich síla a zastoupení se nemění v jednotlivých věkových kategoriích. Jelikož se jedná o replikaci článku a v průběhu výzkumu došlo k nesrovnalostem, byl autorům publikace odeslán email s výčtem problémů při replikaci.

Klíčová slova:

Gender, genderový stereotyp, povolání, word embedding, fastText Skipgram

Title of the Bachelor's Thesis:

Gender stereotypes about occupations in natural language

Abstract:

The bachelor thesis is processed as a replication study of an article about gender stereotypes about occupations in natural language. The main areas of research (i.e. corporas) are children's directed speech, children's and adult speech, and books, and audiovisual media for children and adults. The aim of this study is to detect associations between professions and gender implicitly reflected in the textual sources. The theoretical part describes the concepts related to big data and machine learning, replication and replication crisis, gender, and stereotypes, and the analytical methods used. In the practical part the corporas are analyzed using the fastText Skipgram model. The resulting word embeddings are interpreted using methods of meta-analysis in statistical software R. The result of the work is an empirical evidence that gender stereotypes about occupations appear in natural language and correlate with real labor-office statistics. Exist consistent differences in perceptions of gender stereotypes about professions across corporas, and their strength, and representation do not change across age categories. As this is a replication of the article, and inconsistencies occurred during the research, an email was sent to the authors of the publication listing the problems with replication.

Key words:

Gender, gender stereotype, occupation, word embeddings, fastText Skipgram

Obsah

1	Teoretický základ	9
1.1	Big Data	9
1.2	Machine learning	9
1.2.1	Algoritmus	10
1.2.2	Supervised Machine Learning.....	10
1.2.3	Unsupervised Machine Learning	10
1.3	Textová data.....	11
1.4	Replikace.....	11
1.4.1	Replikační krize	11
1.5	Gender.....	12
1.6	Stereotypy	13
1.6.1	Sociální kategorie a sociální skupiny	13
1.6.2	Kategorizace, stereotypizace.....	13
1.7	Word embeddings (word vektory)	13
1.7.1	Word2vec	14
1.7.2	fastText.....	14
1.7.3	Continuous bag of words model (CBOW).....	14
1.7.4	Skipgram	15
1.7.5	WEAT (Word Embedding Association Test)	15
1.7.6	SC – WEAT (Single Category Word Embedding Test)	16
1.7.7	Kosinová podobnost.....	17
1.8	Teoretický popis použitých metod.....	17
1.8.1	Regresní analýza	17
1.8.2	Binomický test	17
1.8.3	Korelace	18
1.8.4	Kovariance	18
1.9	Literární řešerše	18
1.9.1	Zhodnocení literární řešerše	22
1.10	Výzkumné otázky a hypotézy	22
2	Data	24
2.1	Data – oblasti	24
2.1.1	Child - produced speech = Dětská mluva	24
2.1.2	Child - directed speech = Mluva mířená dětem	24

2.1.3	Child - directed books = Dětské knihy	24
2.1.4	Child - directed audiovisual media = TV dětí	24
2.1.5	Adult - produced speech = Mluva dospělých	25
2.1.6	Adult - directed books = Knihy pro dospělé	25
2.1.7	Adult - directed audiovisual media = TV dospělých	25
2.2	Data – korpusy	25
2.3	Použité soubory v bakalářské práci	26
3	Metody	27
3.1	Proces sběru dat a jejich čištění	27
3.2	Stanovení kategorie skupin a atributů	27
3.3	Volba slov zástupných jednotlivých kategorií skupin	27
3.4	Tvorba word vectorů pomocí fastText	28
3.5	Výpočet WEATu	28
3.6	Analýzy v programu R	30
3.6.1	Nesrovnalosti a nutné úpravy	31
3.6.2	Hlavní korpusy pro výzkum a povolání v nich	31
3.6.3	Regresní analýza	32
3.6.4	Binomické testy	32
3.6.5	Korelace	32
3.6.6	Kovariance	32
4	Výsledky	33
4.1	Celková data	33
4.1.1	Výskyt povolání v korpusech	33
4.1.2	Efekty povolání v korpusech	35
4.2	Popis meta-analýzy	36
4.2.1	Nejsilněji asociovaná povolání s genderem napříč korpusy	36
4.2.2	Zastoupení profesí v jednotlivých intervalech velikosti efektu	37
4.2.3	Nejsilnější genderové stereotypy o povolání v jednotlivých korpusech	40
4.2.4	Meta-regrese napříč korpusy, rozdíly ve věkové kategorii	42
4.2.5	Meta-regrese napříč korpusy, rozdíly v čase získání dat	42
4.2.6	Korelace se skutečnými statistikami (U.S. Bureau of Labor Statistics)	43
4.3	Vyhodnocení hypotéz	45
5	Závěr	48
6	Reference	50

Úvod

Bakalářská práce se zabývá replikační analýzou studie článku „Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words” za pomoci metody fastText Skipgram a WEATu a následně metod meta-analýzy. Publikace vyšla v rámci psychologického žurnálu *Psychological Science* (IF = 5,37; AIS = 3,50) (Clarivate, 2021), pojednává o genderových stereotypech a zaměřuje se přímo na jejich výskyt v komunikačních tocích. Zkoumá, jak moc nás používaný (zakotvený) styl komunikace a jazyk ovlivňuje v přirozených asociacích, konkrétně v tomto případě, zda si muže / ženu podvědomě (tj. implicitně) asociujeme s určitým povoláním (Charlesworth, Yang, Mann, Kurdi, & Banaji, 2021).

V uplynulých letech otřásla výzkumem v sociální psychologii a celkově ve všech společenských vědách tzv. replikační krize. Jedná se o problém, kdy k publikovaným studiím v principu nebyla zveřejňována data a postupy, přičemž v nich docházelo jak k nezamýšleným chybám, tak i k tzv. questionable research practices. Autoři si data také někdy pozměňovali a nikdo to nemohl zkontrolovat. Největší pozornost se tímto směrem začala upínat v druhém desetiletí 21. století v oblasti společenských věd a od té doby se dbá na možnost replikace výzkumu (Vávra, 2015, s. 19). Replikací např. experimentálních ale i empirických studií se zjistilo, že více než polovina těchto výsledků není replikovatelná (tj. velikost a statistickou významnost efektů se nepodařilo replikovat), např. Open Science Collaboration (2015).

Gender bývá také označován za sociální pohlaví a pod tímto pojmem se skrývá soubor charakteristik přiřazovaný mužům či ženě (v dnešní době však lidé hovoří o daleko vyšším počtu genderů). Do jeho vnímání zasahuje společnost, kultura a další vnější vlivy. Je ovlivňován časem, lidé v minulosti vnímali muže či ženu jinak, než je tomu dnes. Například ti, kteří se narodili jako muži, měli větší práva než ženy (ženy nemohly v určitých kulturách volit, vlastnit řidičský průkaz nebo byly na nižší sociální pozici než muži), což se postupně na mnoha místech světa změnilo a stále se mění, dochází k genderové vyrovnanosti. Ovšem není tomu tak ve všech případech, i přes faktické změny, někteří lidé mají stále zafixované historické asociace o tom, jak tomu bylo dříve a povědomí populace o znacích daného genderu mohou být někdy zastaralá a zkreslená. Hovoříme poté o genderových stereotypech (Koldinská, 2010, s. 20).

Podobnou analýzou se v roce 2019 zabývali Schulz & Bahník (2019), kteří zkoumali, jaká adjektiva se v textech anglicky psané beletrie 20. století využívaly ve spojení s muži a ženami. Zjistili, že ženy byly prvně líčeny v pozitivnějším světle, ale v polovině století se tato tendence obrátila a muži byli spojováni s kladnějšími přídavnými jmény. Tato práce ukazuje, stejně jako v původním článku, že genderové stereotypy i přes tvrzení některých lidí existují, a jsou konzistentní napříč různými oblastmi zkoumání. Zároveň potvrzuje, že za pomoci dat a postupů, jenž výzkumný tým podílející se na daném článku zveřejnil, je možné dojít ke stejným výsledkům jako ve studii, a proto jsou publikované závěry průkazné.

1 Teoretický základ

Teoretická část postupně popisuje témata relevantní pro bakalářskou práci. Charakterizuje pojem big data, machine learning a algoritmus. Obsahuje popis pojmu replikace a s tím spojenou replikační krizi, následně se přesouvá k genderu a stereotypům a dále se zabývá použitými metodami analýz. Zpracovává podobné publikace, jež byly v průběhu předchozích let zveřejněny v kapitole literární rešerše a ve finále se přesouvá k určení hypotéz.

1.1 Big Data

Velký a rychlý rozmach technologií, větší provázanost světa, dokumentace informací a otevřený přístup k nim, to vše připravilo pole pro vznik Big dat (v překladu z angličtiny *velkých dat*). Hovoříme o datech, která jsou typická pro svůj velký objem v digitálním světě a díky tomu se stávají vhodným předpokladem pro vznik mnoha analýz. Obecně se pro jejich charakteristiku používá termín 3V (Wessels et al., 2017, s. 57):

1. Volume = objem (patří zde enormní množství informací a dat)
2. Velocity = rychlost (jsou vytvářena nepřetržitě, rychle a v reálném čase)
3. Variety = různorodost (data mohou být strukturovaná, polostrukturovaná a nestrukturovaná)

Mezi další vlastnosti big dat patří jejich organičnost, nejsou vytvářena vyloženě za jedním společným účelem, ale jejich důvodů pro vznik je nespočet. Patří zde data o vědě, přírodě, počasí, data ze sociálních sítí, z map, historická data, recenze a mnoho dalšího (Kitchin, & McArdle, 2016, s. 1–3).

Na základě výzkumu z půdy MIT (Massachusetts Institute of Technology) bylo zjištěno, že na každých 7,7 milionů lidí připadá vyprodukování 1,7 megabytů každou vteřinu každého dne. Zároveň pouze 0,5 % těchto dat je analyzováno a použito. S přibývajícím objemem a dostupností dat přichází i řada potenciálních rizik. Lidé se mohou obávat o zvýšené ohrožení soukromí na internetu, proto je tak důležité porozumět vztahům mezi lidmi a digitálním světem a tím, co se v něm děje (MIT Experimental Learning In Collaboration with GetSmarter, 2016).

1.2 Machine learning

Machine learning (v překladu *strojové učení*) je nástroj patřící do oblasti umělé inteligence, který na základě určitých technik a algoritmů umožňuje počítačovému systému „učit se.“ Každá platforma používající machine learning sbírá co největší množství dat a díky nim je poté schopna předpovědět a vyvozovat závěry. Ty se s každým okamžikem mohou

změnit, jelikož data jsou produkována nepřetržitě a na to strojové učení reaguje (Mullainathan, & Spiess, 2017, s. 89).

Tento systém (= recommender system) známe například z platform jako je YouTube, Netflix nebo Spotify, kdy nám po přehrání jednoho nebo pár videí, filmů či hudby daná internetová stránka či aplikace automaticky nabídne něco podobného, protože je už naučena na fakt, že se nám líbí tento styl (Hao, 2018).

1.2.1 Algoritmus

Za algoritmus považujeme postup při řešení problému. Sestává z logické a řídicí složky, kdy za logickou část považujeme znalosti, které již známe a mají být použity při řešení problému. Zároveň logická komponenta udává význam algoritmu. Jako řídicí složku označujeme strategii řešení daného problému, která pracuje s již zmíněnými znalostmi, navrhuje daný algoritmus a může za jeho účinnost. Při zlepšení efektivity algoritmu dochází ke změně řídicí složky, přičemž ta logická zůstává stejná. Každý algoritmus by měl mít nějaký výstup, který je řešením zadané otázky (Kowalski, 1979, s. 424).

1.2.2 Supervised Machine Learning

Supervised Machine Learning, nebo – li učení s učitelem, je nejčastější typ strojového učení, při kterém stroj ví, co má hledat – ví, která proměnná je target (tj. jednotlivé instance mají svůj label). Na základě vzorku dat, ve kterém jsou pozorované hodnoty target i features se algoritmus naučí vztah mezi těmito proměnnými (v principu jde o odhad funkce, resp. inference pravidla mezi targetem a features). Toto pravidlo se pak dá využít na tzv. out-of-sample predikci, což jednoduše znamená predikci nepozorovaného targetu z pozorovaných features. U učení s učitelem se predikční schopnost algoritmu nutně nezlepší s přidáním více dat – záleží totiž na tom, zdali nová data přináší více signálu o vztahu mezi targetem a features než je v dosavadním vzorku dat k dispozici. Data jsou pro analýzu a vyvození výsledku dostatečně charakterizována a je znám cíl výzkumu.

1.2.3 Unsupervised Machine Learning

Termín Unsupervised Machine Learning můžeme přeložit jako „učení bez učitele“ a jedná se o analýzu dat, která sama o sobě nenesou labely pro rozlišení a kategorizaci (tj. není známa target proměnná a jednotlivé instance nemají svůj label). Úkolem Unsupervised Machine Learning je pak najít jakékoliv vzorce, které by daná data mohla rozřadit do skupin podle podobnosti a našly mezi nimi vztahy (Hao, 2018). Oproti učení s učitelem se algoritmy učení bez učitele vždy zlepši s přidáním více dat, protože nová data zvýší poznávací schopnost algoritmu.

1.3 Textová data

Textová data mívají podobu obyčejného textu, za kterým se skrývá nějaký význam. Nejsou nijak strukturovaná jako například data v tabulkách, což z nich činí složitější vstup pro jakékoliv analýzy (Gentzkow, Kelly, & Taddy, 2019, s. 1).

1.4 Replikace

Replikace je schopnost dojít ke stejným nebo alespoň podobným výsledkům za pomoci stejných metod a stejných dat, jako použil autor. Je tedy potřeba znát postupy studie a mít k dispozici zkoumaný soubor dat. Replikace by měla být beze změn, a to přesně podle předchůdce, který daný výzkum publikoval (Drummond, C., 2017, s. 4).

Pokud se tématem zabývá více lidí a opětovně docházejí k podobným výsledkům, tak dané publikace a výzkumy nabývají na věrohodnosti. Zároveň je stále častější, že významné instituce schopnost replikace výzkumu požadují.

Existuje řada argumentů, proč výzkumy replikovat. Ty ve svém článku shrnul například Chris Drummond. Tvrdí, že schopnost reprodukce vždy byla a je součástí vědy, umožňuje dosáhnout pokroku, protože lidé mohou vycházet ze starších prací a postupů a mohou určité metody přejímat a použít je ve svých výzkumech. Dále je dle něj nutné zabránit vědeckému pochybení a případně ztrátě důvěry ve vědu, k čemuž slouží nutnost předložení údajů a výpočetních nástrojů, které pomohly k dosažení výsledků (Drummond, C., 2017, s. 3).

1.4.1 Replikační krize

Téma replikační krize se stalo za poslední dobu výrazně diskutovaným v akademickém světě. Jedná se o kauzy z druhého desetiletí 21. století, kdy vědci v různých vědních disciplínách publikovali články a v nich udělali chyby nebo si data přizpůsobili. Tato jejich díla nikdo nemohl zkontrolovat, jelikož nezveřejňovali data, ze kterých ve svých výzkumech vycházeli. Kdyby jejich postupy byly v době publikace dostupné, mohlo by dojít k nezávislé kontrole další stranou a mohlo by se předejít případným chybám. Nejčastěji je termín replikační krize spojován s psychologií, ale je zřejmé, že se tento problém vyskytuje prakticky ve všech oborech, kde dochází k analýze dat (Vávra, 2015, s. 19).

Mezi zjištění reálné situace patří výzkum společnosti OSC (Open Science Collaboration) z roku 2015, kdy proběhlo přezkoumání 100 psychologických pokusů o replikaci v projektu Reproducibility Project – Psychology (RPP) a bylo dokázáno, že pouze 36 % vykazovalo statisticky významné výsledky a bylo signifikantních. Dále například farmaceutická společnost Bayer byla schopna zreplikovat pouze 25 % výzkumných studií o rakovině z celkového vzorku 67 studií. Mnoho lidí proto zaujalo postoj, že zjištění nejsou

reálná. Proto se na schopnost replikace v dnešní době tak dbá a je považována za „zlatý standard“ vědy (Anderson, & Maxwell, 2017, s. 1).

Často se implicitně předpokládá, že schopnost replikace ovlivňuje důvěru veřejnosti ve vědu a její výzkumy. Vědci z německé University of Münster zkoumali pohled veřejnosti na publikace a schopnost replikace. Zaznamenávali, zda byla replikace úspěšná nebo nebyla a také se zaměřovali na autora a replikátora publikace. Zaměstnali vzorek lidí (N = 484) ve věku 18–71 let, kteří měli dané oblasti hodnotit. Výsledky ukázaly, že stupeň věrohodnosti článku a důvěryhodnost či stupeň předpokládané odbornosti autora byly silně závislé na tom, zda replikace byla úspěšná či ne. V případě úspěchu byly obě proměnné hodnoceny pozitivněji než v případě neúspěchu. Na replikátora se pohlíželo v obou případech stejně (Hendriks, Kienhues, & Bromme, 2020, s. 1, 5) Znamená to tedy, že pokud je publikace replikovatelná, veřejnost pohlíží na autora i danou práci s větší důvěrou, než pokud se publikace replikovat nedá.

1.5 Gender

Gender někdy bývá označován jako sociální pohlaví. Jsou to vlastně projevy, kulturní charakteristiky a modely, které jsou přiřazovány až vnucovány určitému pohlaví (myšleno z hlediska biologického). Přitom platí, že gender se nerovná pohlaví. V minulosti to však bylo chápáno jinak. Existoval model, který připodobňoval až vnucoval charakteristiky genderu biologickému pohlaví, přičemž nepočítal s odchylkami od předepsaného stereotypu. Vysvětloval, že existují dvě pohlaví a tím pádem pouze dva gendery, které se od sebe silně liší, ale lidé v rámci jedné této kategorie vykazují značnou podobnost (Holmes, Meyerhoff, & Blackwell Reference Online, 2005, s. 23).

Tento názor se v dnešní době netěší velké popularitě. Objevují se lidé, kteří se neidentifikují s daným pohlavím, které mají zapsáno v rodném listě. Tato část populace se označuje za transgenderovou. Už jenom daná předpona “trans” znamená, že se jedná o něco “křížem,” tudíž o něco opačného. Naopak lidé, kteří cítí, že přísluší danému genderu, se nazývají termínem cisgender. Každopádně stále zde vyvstává fakt, že někteří lidé se neidentifikují ani s jedním genderem, binární rozdělení nedokáže správně vystihnout jejich pocity, a proto se necítí komfortně v těchto označeních. Raději používají termíny jako non binární nebo genderově nekomfortní (GNC) (Fiani, & Han, 2018, s. 3).

Gender se v binární soustavě (muž x žena) utváří v průběhu socializace člověka a jedná se o učení se daného jedince, jak být ženou či mužem. V minulosti se totiž vytvořily společenské role a úkoly, které vyplývaly z biologických vlastností daného pohlaví a toto vnímání přetrvalo, ale vnější podmínky se změnily. Někteří lidé ovšem pořád podvědomě očekávají, že historické normy jsou stále platné, a tudíž žena i muž by měli vykazovat znaky typické pro jejich pohlaví. Tato tvrzení nazýváme jako genderové stereotypy (Koldinská, 2010, s. 20).

Genderové stereotypy jsou obecně škodlivé pro skupiny, které jsou v rámci těchto stereotypů vykreslovány jako podřadné. Genderové stereotypy se přitom šíří i prostřednictvím folkloru – souborem tradičních domněnek, zvyků a příběhů, které jsou ústně předávány z generace na generaci. Tuto škodlivost tak čerstvě ukázali i Michalopoulos & Xue (2021), kteří na inovativním datasetu zjistili, že společnosti, které v příbězích a pohádkách vykreslují muže jako dominantní a ženy jako submisivní mají tendenci delegovat ženy do podřadných pozic ve svých komunitách – a to historicky i dnes.

1.6 Stereotypy

Jakožto stereotyp označujeme zjednodušeně řečeno dojem, který v lidech zůstává po určité zkušenosti. Pokud se setkají s někým nebo nějakým jevem, ukládají si to do své paměti a v budoucnu budou v podobných situacích očekávat takové věci, které zažili v minulosti. Pan docent Hnilica uvádí jakožto příklad střetnutí jedince s několika Mongoly, kteří upoutali jeho pozornost svou výškou, jelikož všichni měřili přes 2 metry. Tvrdí, že člověk, který v předchozí době nikdy neviděl příslušníky mongolského národa, si tuto skutečnost zafixuje a pokud se v budoucnu bude znovu hovořit o Mongolech, automaticky si vybaví vysoké jedince (Hnilica, 2010, s. 12–13).

1.6.1 Sociální kategorie a sociální skupiny

S pojmem stereotyp se vážou také sociální skupiny nebo kategorie. Jimi rozumíme uskupení lidí, kteří vykazují podobné znaky v určité oblasti. Pokud vezmeme v úvahu předchozí příklad, tak sociální kategorií by byl mongolský národ.

1.6.2 Kategorizace, stereotypizace

Proces řazení jedinců do jednotlivých sociálních kategorií se nazývá kategorizace. Dochází přitom k observaci jevu, následnému „kódování“ a zařazení do určité kategorie.

Stereotypizace je naproti tomu jev následný. Zahrnuje sociální percepci (utváření dojmu) a kauzální atribuci (vysvětlování pozorovaného chování) (Hnilica, 2010, s. 12–13).

1.7 Word embeddings (word vektory)

Jakožto pojem word embedding rozumíme numerickou reprezentaci slov, stejně jako v různých programovacích jazycích je každé barvě přiděleno určité číslo, každému slovu může být přidělen několikadimenzionální vektor na základě jeho významu. Díky těmto vektorům se následně za pomoci kosinové podobnosti dají vypočítat hodnoty (efekty), které značí o podobnosti slov a měří asociace mezi nimi.

Slova podobného významu budou mít vektory umístěné v prostoru blízko u sebe a směřující podobným směrem, tudíž se vytvoří malý úhel mezi nimi a budou vykazovat i vysokou kosinovou podobnost. Naopak slova významově vzdálená budou mít vektory ukazující na strany opačné, budou v prostoru daleko od sebe, budou mít větší úhel mezi sebou a nebudou si v hodnotě cosinu tak podobné. Pro počítání word vektorů existuje několik metod a v textu bakalářské práce jsou zmíněny dvě: fastText a Word2vec, které pracují na bázi strojového učení. Obě metody mohou také fungovat v podobě dvou modelů: CBOW, Skip Gram. Pro následné zjištění hodnot podobností word embeddingů a tím pádem podobnosti významu slov existuje metoda zvaná Word Embedding Asociation Test (WEAT) (Mikolov, Chen, Corrado, & Dean, 2013).

1.7.1 Word2vec

Model Word2vec (doslova „slova do vektorů“) se stal revoluční metodou pro analýzu textu a byl vynalezen českým počítačovým vědcem Tomášem Mikolovem a spol. Její princip spočívá v tom, že existuje prostor, ve kterém se vyskytují významy slov a každé celé slovo zaujímá svou pozici, jež je vyjádřena vektorem (Rong, 2016, s. 17). A stejně jako u fastTextu, jedná se o doplněk (library) poskytující funkce analýzy textu (Mikolov, Chen, Corrado, & Dean, 2013, s. 1–2).

1.7.2 fastText

FastText je metoda (a zároveň library – doplněk poskytující určité rozšíření funkcí např. v Pythonu) vytvořená Facebookem, která umožňuje efektivní učení slovních reprezentací a klasifikaci vět. Předpokládá, že každé slovo se dá vyjádřit multidimenzionálním vektorem. Tento vektor vzniká na základě průměru vektorů jednotlivých n-gramů. N-gramy jsou jednotlivé části slova, např. slovo „gram“ můžeme rozložit na n-gramy: „<gr“ ; „gra“ ; „gram>“ ; „ra“ ; „ram>“ ; „am>“. Model fastText je v principu rozšířením původního modelu word2vec.

1.7.3 Continuous bag of words model (CBOW)

Predikuje target (slovní reprezentaci) za pomoci kontextu, jako kontext rozumíme zafixované okénko slov okolo, ze kterých se vyvodí výsledný target. Pro srozumitelnost je uvedena příkladná věta:

„Byl krásný den a všichni šli k vodě, jen pár lidí muselo dopisovat bakalářskou práci.“

Jakožto target je zvoleno např. slovo „dopisovat.“ CBOW se podívá v případě fixačního okénka stanoveného na hodnotu 2 na 2 slova před targetem a 2 slova za targetem (lidí, muselo, bakalářskou, práci) a provede součet jejich vektorů, načež předpoví, jak by mohl target vypadat.

Tento předpovězený target má také svůj word vektor, který je porovnán se skutečným původním targetem na základě kosinové podobnosti. Tento proces je zopakován pro každý výskyt daného slova v celkovém textu.

Čím větší objem slov v textu, tím vyšší přesnost při stanovování podobností mezi jednotlivými slovy, při stanovování síly efektů (Mikolov, Chen, Corrado, & Dean, 2013, s. 4).

1.7.4 Skipgram

Skipgram model též funguje na principu predikce okolních slov targetu. Zaměří se na slovo a předpoví, jak by mohla vypadat slova okolo a další postup je stejný jako u CBOW. Předpovězená slova mají také svůj word vektor a ta se porovnávají s vektory skutečných slov v textu. Skipgram to takto dělá pro každé slovo v korpusu, načez výslednou podobností mezi slovy je průměr všech kosinových podobností (efekt) týkajících se daných slov. Pro srozumitelnost použiji stejnou větu:

„Byl krásný den a všichni šli k vodě, jen pár lidí muselo dopisovat bakalářskou práci.“

Jako target je stanoveno slovo „dopisovat.“ Skipgram se zaměří na toto slovo a za pomoci jeho vektoru se snaží předpovědět podobu slov okolo (Mikolov, Chen, Corrado, & Dean, 2013, s. 4).

1.7.5 WEAT (Word Embedding Association Test)

Jedná se o metodu – test, který přemění jednotlivé word vectory na tzv. effect sizes (= velikosti efektů), což je v tomto případě velikost síly genderových stereotypů. Počítá tzv. standardizovanou míru velikosti efektu relativního vztahu mezi slovy představující kategorie skupin (v případě bakalářské práce: muž x žena) a slovy zastupujícími atributy (povolání).

V první fázi se stanoví dvě oblasti zkoumání, do té první patří dvě opozitní kategorie (muž x žena) a do druhé kategorie se zařadí atributy (můžou to být právě povolání nebo jakékoliv jiné cílové skupiny, pro které se bude počítat hodnota asociace s první kategorií – mužem a ženou). Následně je stanovena nulová hypotéza a to, že neexistují rozdíly mezi podobnostmi dvou členů zmíněných dvou oblastí zkoumání (opozitní kategorie a atributy). Dalším krokem jsou permutační testy, které počítají pravděpodobnost (resp. cutoff / threshold) pro (ne)zamítnutí nulové hypotézy. Opakovaně se obměňují atributy, měří průměry hodnot asociací a detekují větší rozdíly, z čehož se stanoví požadovaná pravděpodobnost (Cvencek, Meltzoff, & Greenwald, 2011, s. 1). Rovnice, které hodnoty WEATu počítají, vypadají takto:

$$\text{sim}(X, Y, A, B) = \sum_{x \in X} \text{sim}(x, A, B) - \sum_{y \in Y} \text{sim}(y, A, B)$$

X, Y – zastupují dva členy opozitní kategorie (muž x žena)

A, B – zastupují dva členy atributů (dvě povolání, např. mechanik, zdravotní sestra)

$$\text{sim}(w, A, B) = \phi_{a \in A} \cos(\vec{w}, \vec{a}) - \phi_{b \in B} \cos(\vec{w}, \vec{b})$$

ϕ – označuje průměrnou hodnotu

$\cos(\vec{a}, \vec{b})$ – označuje úhel kosinu mezi vektory \vec{a} a \vec{b}

$\text{sim}(w, A, B)$ – měří asociaci atributu s w

$\text{sim}(X, Y, A, B)$ – měří diferencovanou asociaci atributů s členy opozitní kategorie

Pro permutační testy je následně stanovena rovnice, která počítá s jednostrannou p – hodnotou.

$$\text{Pr}_i [\text{sim}(X_i, Y_i, A, B) > \text{sim}(X, Y, A, B)]$$

Pr_i – označuje jednotlivý permutační test

$\{(X_i, Y_i)\}$ – rozděljuje všechny segmenty $X \cup Y$ do dvou částí stejné velikosti

$X \cup Y$ – označení sjednocení množiny X a Y

A finální rovnice pro výpočet hodnoty WEATu (asociací) zní:

$$\frac{\phi_{x \in X} \text{sim}(x, A, B) - \phi_{y \in Y} \text{sim}(y, A, B)}{\sigma_{w \in X \cup Y} \text{sim}(w, A, B)}$$

σ – označuje směrodatnou odchylku

1.7.6 SC – WEAT (Single Category Word Embedding Test)

SC – WEAT je jedno – kategoričtý WEAT, počítá se tedy pouze s jednou kategorií a nedochází v něm k double difference efektu. WEAT počítá s oběma kategoriemi a pokud zkoumá například vztah, který je zmíněn v studii č. 1 replikovaného článku, tedy muž – práce a žena – práce, dojde následně při zkoumání vztahu muž – domov a žena – domov k podobnému výsledku, a právě zde dochází k double difference efektu. Protože pokud bereme v úvahu, že existuje stereotyp „muž – práce,“ tak je to prakticky to samé jako „žena – domov“ (Charlesworth et al., 2021, s. 224).

1.7.7 Kosinová podobnost

Měří podobnost mezi 2 vektory v určitém prostoru a je definována vzorcem (Han, Kamber, & Pei, 2012, s. 77–78):

$$\text{sim}(x, y) = \frac{x * y}{\|x\| * \|y\|}$$

$\|x\|$ označuje tzv. Euklidovskou normu vektoru a pro $x = (x_1, x_2, \dots, x_p)$ platí:

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$$

Pozn.: Euklidovská norma vektoru je rovna euklidovské vzdálenosti počátečního a koncového bodu libovolného umístění tohoto vektoru. Euklidovská vzdálenost je rovna délce úsečky, která tyto body spojuje (v dvou-dimenzionálním prostoru).

1.8 Teoretický popis použitých metod

V následující kapitole jsou popsány, metody, kterými jsou data analyzována. Jedná se o set metod meta-analýzy a v této práci se jedná o regresní analýzu, binomické testy, korelace a kovariance.

1.8.1 Regresní analýza

Regrese zjišťuje, jak aspekty podílející se na výsledku daného předmětu zkoumání, ovlivňují jeho různorodost. Regresní metody v principu odhadují funkci mezi targetem a prediktory, kde odhady jednotlivých parametrů regresního modelu ukazují směr, velikost a statistickou signifikanci efektu prediktoru na target (v mikroekonometrické terminologii se target také nazývá závislou / response proměnnou či řídčejí regresandem), pomocí regrese se snažíme odhadnout hodnotu náhodné veličiny (závislé proměnné) na základě poznatků o určité jiné veličině (nezávislé proměnné).

Důvodem pro využití regresní analýzy je typicky snaha zjistit, jak velkou váhou každý aspekt dat přispěl k vytvoření výsledku, měří se tedy tzv. „effect sizes“ (Tatsioni, & Ioannidis, 2017, s. 120).

1.8.2 Binomický test

Binomický test ověřuje, zda procentuální výskyt určitého zkoumaného jevu, jedné ze dvou kategorií, odpovídá očekávané frekvenci (dle binomické pravděpodobností distribuce).

Jedná se o přesný test statistické významnosti odchylek od očekávaného rozdělení pozorování kategorií, tj. test dává skutečně „exact p – values“.

1.8.3 Korelace

Termín korelace se používá pro sílu vztahu-vzájemné závislosti mezi dvěma veličinami. Jedná se o statistický lineární vztah a říká, když se jedna veličina změní, změní se i druhá. Korelace je vyjádření míry souvislosti mezi těmito dvěma veličinami a zároveň se může použít v případě jakýchkoliv jednotek veličin. Je vyjádřena korelačním koeficientem, který nabývá hodnoty v intervalu $[-1; 1]$ a pokud je hodnota korelace 1, znamená to 100% pozitivní souvislost. V případě hodnoty korelace -1, hovoříme o negativní souvislosti a při hodnotě 0 se jedná o žádnou souvislost mezi veličinami. Korelace samozřejmě neznamená kauzalitu, klasická frekventistická statistika si nedělá žádné nároky na to, zdali x ovlivňuje y, nebo je to naopak. Tím se zabývají metody kauzální inference, které budou podle mého názoru hrát v dalším vývoji strojového učení ještě velkou roli a postupem času nakonec dojde k vytvoření tzv. causal machine learning oboru. Zde bych jen pro čtenáře poznamenala, že ačkoliv korelace automaticky neznamená kauzalitu, některé korelace ale kauzalitu znamenají. Záleží na tom, zdali je v mechanismu přítomen confounder či ne.

1.8.4 Kovariance

Kovariance také vyjadřuje vzájemný vztah dvou spojitých veličin, které jsou vyjádřeny stejnými jednotkami. Počítá s rozptyly veličin a jedná se o společnou hodnotu rozptylu. Stejně jako u korelace, kovariance nabývá hodnot v intervalu $[-1; 1]$ a pokud hodnota kovariance nabývá hodnot větších než 0, jedná se o pozitivní souvislost, pokud je menší než 0, hovoříme o negativní souvislosti a v hodnotě 0 se jedná o žádnou souvislost mezi veličinami.

1.9 Literární rešerše

To, že genderové stereotypy nacházející se v přirozeném jazyce mohou ovlivňovat počet lidí na různých pracovních pozicích, zkoumali již Lewis a Lupyan (2019) ve své práci „Gender stereotypes are reflected in the distributional structure of 25 languages.“ Konkrétně se zabývali verifikací kulturních stereotypů jakožto myšlenky, že žena by měla spíše zůstat doma s dětmi, kdežto muž by měl chodit do práce a postarat se o rodinu finančně. Měřili genderové asociace zakotvené v 25 jazycích světa a zároveň analyzovali, zda lingvistické asociace utvářejí implicitní úsudky lidí.

Tato publikace používá projekt Project Implicit, který funguje na bázi Implicit Association Testu a vznikl v roce 1998 za spolupráce 3 vědců – Mahzarin Banaji z Harvard University, Anthony Greenwald z University of Washington a Brian Nosek z University of Virginia. IAT měří implicitní genderové asociace, tedy ty, které jsou podvědomé a člověk si je nemusí sám od sebe uvědomit. Projekt Implicit nabízí testy z různých oblastí, měří například implicitní asociace týkající se rasových témat, násilí, nicméně pro tuto práci jsou relevantní právě testy týkající se genderu (Project Implicit, 2011).

Test účastníkovi ukáže dvě základní opozitní kategorie, ke kterým následně bude muset přiřazovat slova, jež do dané kategorie spadají. Například kategorií bude pohlaví, tudíž muž a žena a na obrazovce se budou objevovat pojmy jako dcera, dědeček, matka, on, a úkolem účastníka bude co nejrychleji přiřadit pojem do správné kategorie. Následně přibude další kategorie, pro kterou jsou stereotypy zkoumány. Bude se jednat například o dvojici kariéra a domov, tudíž se účastníkovi budou zobrazovat pojmy jako práce, dům, peníze, děti a úkolem je opět co nejrychleji spojit slova, která k sobě patří. V dalším kroku se prvky těchto dvou kategorií zobrazí pod sebou, půjde tedy na jedné straně o spojení muž – kariéra a na druhé žena – domov a opět se budou ukazovat slova jako dům (účastník musí pojem spojit se stranou, kde se vyskytuje domov). V posledním kroku se prvky kategorií obmění a půjde o spojení žena – kariéra a muž – domov a opět účastník spojuje pojmy se stranou, kam pojem patří. Každá kategorie a každý pojem, který do ní patří, je ještě pro větší přehlednost zobrazen stejnou barvou. Konečné výsledky se vyhodnotí podle rychlosti a správnosti účastníka a podle toho lze dle Projectu Implicit zjistit, jestli podvědomě tíhneme k tvrzení, že muž by měl budovat kariéru a žena měla zůstat doma nebo naopak (Project Implicit, 2011).

Caliskan, Bryson a Narayanan (2017) se zaměřili na stereotypy detekované díky již zmíněnému Implicit Association Testu a za pomoci machine learningu analyzovali různé korpusy. Dospěli k zjištění, že vydávané texty obsahují zakotvené stereotypy z dob historických. Stereotypy byly různé, mohly být morálně neutrální, kdy se například týkaly témat jako květiny a hmyz, nebo se dotýkaly témat genderu a vlastností s nimi spojených, či pracovních preferencí. Hlavně se zabývali na nejčastěji se vyskytující stereotypy, že muži jsou spojováni s kariérou a vědou a ženy se zůstáním doma a s uměním, dále také zkoumali asociace jednotlivých povolání s daným genderem, které porovnávali s reálnými statistikami. Dalším předmětem zkoumání bylo zjištění asociací mezi zbraněmi a charakteristikou „špatné“ a hudebními nástroji a charakteristikou „dobré“ a naopak. Zároveň se zabývali reakcemi na Euro – Americká a Afro – Americká jména a zda jsou spojováni s kladnými či negativními popisy. Pro svůj výzkum vynalezli jinou metodu než IAT, a to WEAT (Word Embedding Association Test). Své hypotézy, že stereotypy o těchto tématech existují a jsou detekovatelné v textu, potvrdili a s tím i výsledky patrné v IATu. Zdůrazňovali fakt, že word embeddingy dokážou poznat sémantický význam jednotlivých spojení, ač nemají zkušenost s reálným světem.

Cvencek, Melzoff a Greenwald (2011) se zabývali stereotypy v okruhu školáků prvního stupně základní školy ohledně matematiky, která lze následně asociovat s vědou. Děti tedy vyplňovaly Implicit Association Test, který byl původně používán u dospělých a je popsán výše v textu bakalářské práce, ale autoři publikace ho upravili, aby byl pro školáky jednodušší a relevantní vůči matematice. Jakožto druhou opozitní kategorii vůči matematice zvolili čtení, v testu byly zároveň zahrnuty otázky týkající se pohledu na jejich vztah s matematikou a čtením, aby se zahrnula i rovina subjektivní. Už na základě předchozích studií byla v práci zahrnuta hypotéza, že “matematika je pro kluky” a to se autoři snažili obhájit nebo vyvrátit. Zjistili, že už od útlého věku tyto asociace vznikají na úrovni implicitní i explicitní. Holčičky častěji volily možnost, že mají raději čtení než matematiku a kluci tato tvrzení nijak nevyvraceli. Podvědomé reakce obou skupin prakticky kopírovaly výsledky testů subjektivních a potvrdily, že s mužským pohlavím je častěji spojována matematika i na prvním stupni základní školy.

Téma profesní segregace ve třech různých desetiletích (90. léta 20. století až 10. léta 21. století) zkoumala americká psycholožka Beyer (2016). Za pomoci tří studií zkoumala povědomí o rozdělení zaměstnanosti v profesích a odměňování mužů a žen za vykonanou práci v USA. Pro zjištění hodnoty průměrného příjmu použila statistiky BLS (U.S. Bureau of Labor Statistics), které popsaly průměrný příjem za rok 2014 \$59,450 pro muže a \$47,076 pro ženy. Zároveň ale zmiňuje, že Američané si myslí, že dosáhli genderové vyrovnanosti a diskriminace ustala. Výzkum proběhl mezi univerzitními studenty a zahrnoval otázky týkající se povědomí o rozdílech mezi muži a ženami v oblasti práce a poslední studie také zjišťovala vztah jedinců k modernímu sexismu. Výsledky odhalily, že studenti si byli vědomi profesní segregace a rozdílů v příjmech, nicméně vysoce podceňovali hodnotu rozdílů. Výsledky byly prakticky konstantní za všechna desetiletí, a proto se autorka domnívá, že by tato situace mohla bez obtíží pokračovat.

Garg, Schiebinger, Jurafsky a Zou (2018) popisovali ve svém výzkumu, jak se měnili stereotypy o ženách a etnických menšinách v průběhu 20. a 21. století v USA. Pomocí word embeddingů zachytili změnu vnímání daných subjektů v různých časových obdobích, kdy probíhaly významné celosvětové akce narážející na společenské poměry (např. hnutí žen v 60. letech 20. století a asijská imigrace do Spojených států amerických). Zjistili, že v průběhu desítek let se adjektiva asociována s ženami vcelku změnila, ale hlavně se zvýšila síla daných asociací. V dřívějších letech byla přídavná jména nejsilněji spojována s muži, ale postupem času nabývají na síle asociace s ženami. Autoři odhadují, že krátce po roce 2020 dojde k vyrovnání síly efektů asociací s oběma gendery. Dále bylo zjištěno, že adjektiva spojována s asijskými jedinci se silně změnila a přelomem byl rok 1950. V letech předchozích s nimi byla silně spojována negativní přídavná jména a po roce 1950 a hlavně 1980 se trend změnil a Asiaté začínali být vykreslováni jako např. sensitivní, proaktivní a vřelí. A v oblasti povolání potvrdili, že implicitní asociace v textech korelují se skutečnými administrativními statistikami.

Podobným tématem jako je téma této bakalářské práce se již zabývali Schulz a Bahník (2019). Zkoumali adjektiva, která si lidé asociují s mužským či ženským pohlavím. Předmětem zkoumání byla anglická beletrie dvacátého století a data byla získána za pomoci Google Books Ngram corpus. Pro použití přídavných jmen použili již vytvořený list s 555 nejpoužívanějšími adjektivy z roku 1968, který nakonec zkrátili na 155 nejčastěji používaných. Přídavná jména byla následně ohodnocena váhami v rozmezí 0–6 (0 nejnižší, 6 – nejvyšší) v oblastech jako femininity x masculinity, incompetence x competence, coldness x warmth, undesirable x desirable. Cílem bylo zjistit vykreslení žen (dívek) a mužů (chlapců) v anglickém jazyce dvacátého století a zjistit, jak se na daná pohlaví pohlíželo, co s nimi bylo spojováno.

Dle výsledků bylo každé pohlaví popisováno pozitivně, nicméně v první polovině dvacátého století se jednalo o kladnější popis žen, v polovině století došlo ke zlomu a pozitivnější adjektiva se čím dál častěji spojovala s muži. Dále byli muži častěji spojováni s výrazy jako „moudrý, schopný, upřímný, líný a lakomý“ a ženy zase s „okouzlující, žárlivá, sprostá, pošetilá.“

Měřením rozdílů v přístupech publika k jednotlivým pohlavím jakožto prezentujícím v rámci ekonomických seminářů, konferencí a jiných setkání se zabývali Dupas, Modestino, Niederle a Wolfers (2021). Dobrovolná skupina 97 kóderů a kóderek, kteří se zabývají genderovou rovností, se postupně na jaře, v létě a v zimě zúčastnila 462 událostí s ekonomickým tématem a mapovala všechna vzájemná působení mezi publikem a prezentujícím mužem či ženou. Zaznamenávali například čas, trvání, typ a smysl každé interakce. Dále zapisovali četnost otázek a vlastnosti každého zapojeného účastníka, věci jako např. jeho tón a zda se jednalo o kritiku nebo pouhý dotaz. Zjistili, že v oblasti zacházení s mužským a ženským prezentujícím se vyskytují významné rozdíly. Ženy jsou například častěji dotazovány, a to až o 12 %. Vysvětlují to tím, že ženy většinou přitahují rozmanitější publikum, proto počet otázek může vzrůst. Dále byl také u tázajících na událostech žen častěji detekován nepřátelský nebo patronující tón. Výzkum byl prováděn kvantitativně i kvalitativně a sesbíraná data byla analyzována za pomoci regrese. Celkové výsledky mohou vysvětlovat, proč se na vyšších manažerských a celkově vedoucích pozicích vyskytují zatím hlavně muži a ženy s podobnými či stejnými kompetencemi na dané místo nedosáhnou.

V oblasti numerických zastoupení slov provedli revoluci Mikolov, Chen, Corrado a Dean (2013), kdy představili nové modelové architektury pro výpočet spojitých vektorových reprezentací slov ve velmi početném datovém prostředí. Své metody porovnávali s dříve dostupnými a dokázali zvýšit přesnost reprezentací slov při mnohem nižších výpočetních nákladech a v kratším čase. Pro rok 2013 poskytli nejmodernější postupy s nejkvalitnějšími výkony pro měření podobností významů slov. Kritizovali modely, které analyzovali slova jakožto atomy a tvrdili, že tímto způsobem nelze najít mezi slovy podobnosti. Jejich metoda proto pojímá slovo a význam slova jakožto nerozdělitelný celek. Cílem práce bylo představit model, který dokáže zanalyzovat datasety čítající biliony slov, jelikož v dané době stávající

modely dokázaly zkoumat „pouze“ několik stovek milionů slov. Zároveň do svého výzkumu implementovali fakt, že existuje několik stupňů podobnosti (např. slova, která mají stejná zakončení). Pro vytvoření nové metody Word2vec (kapitola Word2vec) zkoumali již dříve známé modely Skipgram a CBOW. Word2vec nakonec po analýze vysoce početných datasetů dokázal přiřadit např. slovo Paris k France, když znal rovnici, že Berlin patří k Germany. V závěru práce předpovídali, že v budoucnu by se mohly analyzovat i datasety o velikosti trilionu slov s prakticky jakoukoliv délkou slova.

1.9.1 Zhodnocení literární rešerše

V literární rešerši jsou rozebrány práce relevantní svým obsahem k bakalářské práci. V dnešní době je analýza textu za pomoci machine learning a konkrétně word embeddingů vcelku populárním tématem. Vědci zkoumají podobnosti významů slov a na základě toho se snaží detekovat implicitně zakotvené asociace – stereotypy. Numerická reprezentace slov dokáže zachytit sémantiku slov, která se nachází v textech přirozeného jazyka. Dané analýzy potom měří sílu asociací, početnost a jejich změny v průběhu časových období.

Ve zmíněných pracích je hlavně ukázáno, že genderové stereotypy i např. etnické stereotypy existují a některé se mění, jiné zůstávají konzistentní. V různých kulturách ve světě je tvrzeno, že dochází ke genderové a etnické vyrovnanosti, že neexistují preference a privilegia, a že „všichni lidé jsou si rovni.“ Je potom zajímavé se podívat na pár podobných publikací, které prakticky ve všech případech docházejí k podobným výsledkům. Mezi nejčastěji zkoumané stereotypy patří vztah muže a ženy a kariéry a domovem a vědou a např. uměním. Všechny zmíněné články se ztotožňují s faktem, že v textech přirozeného jazyka jsou muži častěji asociováni s kariérou a vědou (např. matematikou) a ženy se spojují s domovem a uměním (nebo čtením). Většina publikací zároveň prokázala pozitivní korelace se skutečnými administrativními statistikami, proto je možno říct, že stereotypy odráží realitu. Každopádně jak už bylo popsáno v kapitole Gender a Stereotypy, stereotypy jsou škodlivé a můžou ovlivňovat percepci jedinců. Zde vyvstává jen otázka, zda se s tím jedinci či společnost spokojí nebo budou na tato fakta reagovat a dojde ke změně.

1.10 Výzkumné otázky a hypotézy

Předmětem zkoumání je, jak se projevují genderové stereotypy o povolání v komunikaci (textu) a konkrétně jak silné asociace jsou mezi jednotlivými povoláními a mužským či ženským pohlavím. Zájem je také upřen na změny asociací z hlediska času a rozdílných věkových kategorií.

Jakožto první hypotézu stanovíme:

H1: Existují konzistentní rozdíly ve vnímání genderových stereotypů o povolání u mužů i žen v jednotlivých oblastech (napříč korpusy).

Druhou hypotézu formulujeme:

H2: V komunikaci rozdělené podle věkové kategorie se vyskytují podobně silné genderové stereotypy o povolání a také v podobném zastoupení.

Třetí hypotézou bude:

H3: Genderové stereotypy o povolání v přirozeném jazyce korelují s administrativními reálnými statistikami o zaměstnanosti mužů a žen v jednotlivých povoláních.

Zařadíme sem i čtvrtou přídavnou hypotézu, která se týká technické stránky:

H4: Na základě dostupných dat zkoumaného článku lze danou studii replikovat a dojít ke stejným výsledkům.

2 Data

Veškerá data, se kterými se v bakalářské práci pracuje, již použil a zveřejnil tým autorů replikované publikace. Jelikož je jejich výzkum replikován, jsou potřebná ta úplně stejná data, která jsou dostupná na serveru OSF (Charlesworth et al., 2020).

Ústředním záměrem je analýza textu (slov) v komunikačních tocích. Člověk přijímá informace hlavně na základě rozhovorů, z knih nebo z audiovizuálních médií, a proto je upřena pozornost právě na tyto oblasti.

2.1 Data – oblasti

Data pro výzkum jsou získávána v 7 oblastech, které jsou zachyceny ve 12 korpusech:

2.1.1 Child - produced speech = Dětská mluva

Jedná se o data získaná z mluveného projevu dětí. Za pomoci projektu TalkBank, který organizoval Brian MacWhinney z Carnegie Mellon University, byla zmapována mluva dětí a následně přepsána. Jsou to rozhovory z let 1970 až 1990 a čítají přesně 6 518 konverzací tvořených z 8 429 128 slov. Průměrný věk dětí byl 2,92 roku a pohyboval se mezi narozením až dvanáctým rokem života.

Autoři se shodují na tom, že data mohou být trochu zastaralá, nicméně tento korpus je dosud největší v oblasti dětského projevu a stejně velice dobře slouží pro porozumění komunikace mezi rodičem (opatrovníkem) a dítětem.

2.1.2 Child - directed speech = Mluva mířená dětem

Tato data zase pochází ze vztahu rodič (nebo opatrovník) a dítě. Jedná se tedy o přepisy rozhovorů mířených dětem. Data jsou získána ze stejného korpusu jako předchozí kategorie.

2.1.3 Child - directed books = Dětské knihy

V této kategorii se analyzují slova z knih určených pro děti. Konkrétně se využila webová stránka Project Gutenberg, což je doména, kde se nachází více než 60 000 knih v online podobě zdarma. Zkoumaný vzorek obsahuje 98 knih a 4 583 629 slov.

2.1.4 Child - directed audiovisual media = TV děti

Jako zástupci pro audiovizuální média, se kterými děti přicházejí do styku, byla vybrána produkce Disney filmů, Nickelodeon TV Show a PBS Kids TV Show. Jedná se o korpus s

přepisy rozhovorů z 1078 filmů a 4309 televizních epizod. Data byla vybrána z rozmezí let 1938 až po současnost.

2.1.5 Adult - produced speech = Mluva dospělých

Přepisy rozhovorů dospělých lidí jsou získány z korpusu Switchboard - 1 Telephone Speech Corpus. Tvoří ho přibližně 2400 konverzací týkajících se 70 náhodně vybraných témat. Skládají se z 3 063 280 slov a účinkovalo v nich 543 lidí ve věku 17–63 let, přičemž vše bylo mapováno v letech 1990 a 1991. Data byla vybrána díky podobnosti objemu korpusu a době sběru informací jako u kategorie child produced / directed speech.

2.1.6 Adult - directed books = Knihy pro dospělé

Pro výběr zástupců této kategorie byl znovu využit Project Gutenberg a zkoumaný vzorek čítá 1000 náhodně vybraných knih pro dospělé, ve kterých se nachází 40 252 700 slov.

2.1.7 Adult - directed audiovisual media = TV dospělých

Data o audiovizuálních médiích, které se týkají dospělých lidí, byla získána přepisem rozhovorů z filmů a televizních pořadů vytvořených mezi lety 1960 až po současnost. Korpus tvoří 2 056 384 slov (Charlesworth, et al., 2021, s. 3–4).

2.2 Data – korpusy

Každá zmíněná oblast je zachycena ve svém korpusu, který obsahuje slova modifikována do vhodného stavu pro následné analýzy. Většinou je to 1 oblast = 1 korpus, ale v případě TV dětí jsou používány korpusy 4 – TV dětí Nickelodeon, TV dětí Disney, TV dětí PBS TV Show a TV dětí kombinace. Do oblasti TV dospělých patří 2 korpusy – TV dospělých Simply Scripts, TV dospělých PBS TV Show. Mluva týkající se dětí je rozdělena na dvě oblasti, které taky tvoří své vlastní korpusy – Mluva dětí a Mluva mířená dětem, nicméně existuje i kombinace těchto dvou korpusů v jednom obsahující oboje. Obecně můžeme říct, že vždycky existují jednotlivé korpusy a pak existuje jeden, který je v dané oblasti shromažďuje. Co se velikosti týče, největším korpusy jsou Gutenberg a TV dětí kombinace.

Všechny tyto korpusy společně procházejí následnými analýzami, nicméně pouze některé z nich se používají v konečné interpretaci výsledků. Ty, se kterými se v konečné fázi pracuje, stanovili autoři původní práce na: Mluvu dětí a Mluvu mířenou dětem, Mluvu dospělých, Gutenberg, Dětské knihy, TV dětí – kombinace a TV dospělých Simply Scripts.

2.3 Použité soubory v bakalářské práci

Vzhledem k početnosti a velikosti souborů, jenž byly použity pro tvorbu výzkumu bakalářské práce, bylo dle mého názoru a po konzultaci s vedoucím práce vhodnější vytvořit složku na Google Discu, kde jsou všechny soubory rozdělené dle fáze výzkumu práce. Níže je přiložen hypertextový odkaz, který čtenáře do složky přesměruje:

https://drive.google.com/drive/folders/1M3wzp0wlqbGEr3CkWK-0_0iZlcerHvu?usp=sharing

3 Metody

V první části je pozornost upřena na analýzu textu za pomoci metody Skip gram od fastText, následně se efekty počítají pomocí WEATu a na dané výsledky jsou v druhé části aplikovány binomické testy, regrese, korelace a kovariance. V této kapitole je popsáno, jak se dospělo k výsledkům. Analýza je rozebrána od procesu získání dat až po interpretaci výsledků a konečné zhodnocení.

3.1 Proces sběru dat a jejich čištění

Data byla získána z internetové stránky OSF, kde můžeme nalézt veškeré pokyny pro replikaci. Proběhly zde pokusy o samostatné stahování a čištění, ale bohužel se vyskytlo množství chyb a nesrovnalostí, které neumožnily provést tento krok, jako tomu bylo u původních výzkumníků. Proto, a ještě za účelem dosažení co nejpodobnějších výsledků vzhledem k replikaci, byla použita přesně ta data, která jsou na OSF zveřejněna.

Původní tým výzkumníků data získal z materiálů dostupných na internetu a upravil je do podoby textového dokumentu. Zároveň věty připravil o interpunkci a velká písmena a také slova zlematizoval, což znamená, že každé slovo v jiném než původním tvaru, přetvořil na tvar základní (tj. na slovníkový tvar slova), aby se následné analýzy mohly provést bez velkých odchylek. Více informací o tom, odkud přesně jednotlivá data jsou, je napsáno v kapitole Data.

V tomto kroku bylo zjištěno, že korpus, který měl obsahovat slova z audiovizuálních médií pro děti, konkrétně od Disneyho, je nulový. Tato chyba bude zmíněna pomocí emailu tvůrcům studie. Díky zveřejněnému dokumentu se samotnými word vectory pro tento korpus na serveru OSF se dalo pokračovat i bez tohoto základního neopracovaného dokumentu a následné provedení analýzy nebylo poškozeno.

3.2 Stanovení kategorie skupin a atributů

Jelikož se jedná o replikační analýzu, je tento krok podmíněn výběrem tvůrců studie. Vybraná kategorie skupin je muž x žena a atributy je myšleno výčet povolání, pro které se budou jednotlivé efekty počítat.

3.3 Volba slov zástupných jednotlivých kategorií skupin

Pro kategorii skupin je důležité, aby nebyla zastoupena pouze slovy muž a žena, protože tyto pojmy se dají vyjádřit i jinými významově podobnými slovy.

- Pro ženu (female) to je: she, her, mommy, mom, girl, mother, lady sister, momma, mama, sis, grandma, herself
- Pro muže (male) to je: he, his, daddy, dad, boy, father, guy, brother, dada, papa, bro, grandpa, himself

Aby se zjistila tato nejčastěji používaná slova skoro podobného významu, byl původně analyzován korpus s mluvou dětí, kde se autoři zaměřili na frekvenci výskytu daných slov, komplexnost a přesnost vyvážení. Takto získaná slova byla následně zakotvena do dokumentů pythonu, aby se jednotlivé word vectory generovaly vždy i k těmto zástupcům.

3.4 Tvorba word vectorů pomocí fastText

Pro vytvoření word vectorů bylo potřeba stáhnout package pythonu fastText. V původním zadání jsou příkazy v pythonu a příkazovém řádku určeny pro uživatele elektroniky Apple, a jelikož já používám Windows, musela jsem doinstalovat ještě fastText-win, který je prakticky stejný jako originální fastText, ale je určen i pro Windows. Pro tvorbu word vectorů je nadále používán Jupyter Notebook z platformy Anaconda, do kterého se zapíše krátký kód, jež zároveň zmíní, z jakého korpusu se budou vektory extrahovat.

Po dokončení procesu se vytvoří 2 soubory, jeden s příponou .bin a druhý s příponou .vec. Po otevření lze spatřit, že v souborech .vec, se objevují vygenerované multidimenzionální word vectory pro jednotlivá slova z daného korpusu, přičemž na jednom řádku jsou vektory pro jedno slovo. A právě tyto soubory .vec byly pro následné kroky použity. Soubory .bin obsahují celý fastText model.

Vzniklo celkově 22 takových dokumentů (11x .bin a 11x .vec). Díky chybě v korpusu pro dětská audiovizuální média společnosti Disney byl dokument s vektory přejat ze serveru OSF. Takto vytvořené soubory (a 1 přejatý) s vektory pro 12 dokumentů se musel spojit do jednoho dokumentu a přesně pro tento krok se použil skript Pythonu, který sloučil jednotlivé dokumenty s vektory do jednoho za pomoci SQL.

3.5 Výpočet WEATu

1. Výpočet velikosti asociace mezi kategorií skupin a atributu
 - V tomto kroku se počítala kosinova podobnost pro word vector udělen ženě a jejím zástupným slovům a každému word vectoru povolání ze seznamu. To samé se provedlo i u muže a jeho zástupných slov.
2. Průměr asociací mezi kategorií skupin a atributu
 - Následně se jednotlivé kosinovy podobnosti pro word vector ženy a zástupných slov a word vector povolání sečetly a zprůměrovaly, aby vznikla jedna hodnota

reprezentující asociaci mezi ženou a daným povoláním. Totéž se udělalo pro word vector muže a jeho zástupných slov a word vector stejného povolání.

3. Výpočet rozdílu asociací mezi kategorií skupin a atributu

- A nakonec se vezme kosinova podobnost mezi word vectorem muž a word vectorem povolání (například pilot) a od toho se odečte kosinova podobnost mezi word vectorem žena a word vectorem povolání (opět pilot). Takto se to provede pro všechna zkoumaná povolání.

Pro vytvoření podkladu pro analýzu v programu R a následné vyvození výsledků je nutno použít soubor se všemi vektory a umístit ho do složky se souborem obsahující kód, který analýzu provede. Soubor s kódem je získán z úložiště OSF původní práce. Dále je použit příkazový řádek a je nasměrován právě do zmiňované složky, kde se pomocí nadefinovaného souboru pythonu spustí analýza dokumentu se všemi vektory. Výsledkem je excelový soubor se všemi vztahy mezi jednotlivými kategoriemi skupin a atributy.

A jelikož je práce psaná na téma genderové stereotypy o povolání, je nutné provést ještě analýzu pro jednotlivé profese. Pracovní pozice, pro které se dané efekty budou měřit, jsou vypsané v samostatném textovém dokumentu. Postup je prakticky podobný tomu v předchozím odstavci. Je nutné mít tento textový soubor s povoláními ve stejné složce jako excelový soubor z předchozího kroku. Dále je použit kód napsaný v pythonu zaměřující se na analýzu genderových stereotypů o povolání, který vytvořili tvůrci replikovaného článku. Také musí být ve složce soubor se všemi word vectory, jednoduše řečeno je používána stále stejná složka. Následně je pozornost opět věnována příkazovému řádku, který se musí nasměrovat do zmiňované složky se všemi soubory. Zde se poručí pythonu, ať začne analyzovat soubor pythonu s kódem. Výsledkem je opět excelový soubor, který se bude finálně používat při regresní analýze. Tento excelový soubor obsahuje prvky jako excelový soubor předchozí, ovšem s jinými hodnotami a zaměřením na profese.

Konkrétně obsahuje:

a. Korpus

- Vymezení oblasti, ve které jsou efekty počítány (zda se jedná o dětské knihy, média pro dospělé, projev dospělých, ...). Přičemž každý korpus je už upraven do tvaru vhodného pro analýzu (slova jsou v 1. pádu, odstraněna interpunkce,

b. Kategorie skupin

- Vymezení skupin, pro které jsou efekty počítány, v tomto případě se jedná o vztah muže a ženy a je to v každém pozorování stejné, protože se počítá se stereotypy genderu, a v této práci se předpokládá, že existují pouze 2 základní gendery (pohlaví).

c. Atributy

- Jednotlivé prvky, pro které jsou efekty počítány, což v tomto případě je zkoumané povolání a je proměnlivé. Jednotlivé profese se přejímají z textového dokumentu obsahujícího jejich výpis.

d. Velikost efektu

- Velikost efektu, síla asociace atributu s prvkem kategorie skupin. Jaká je standardizovaná míra velikosti efektu relativní asociace mezi slovy představujícími kategorie skupin. V tomto případě síla asociace mezi mužem či ženou a daným povoláním.

e. Chybějící slova pro daný korpus a test

- Jaká slova nebyla použita, ale v ostatních pozorováních ano. Například v jednom pozorování pro prvek kategorie skupiny muž chybí jeho vyjádření jakožto „dada, momma.“

f. p – hodnoty

- p – hodnoty jsou vypočítané z permutačních testů pro velikosti efektů WEAT a ve statistice tento pojem znamená nejmenší hladinu významnosti testu, na které je nulová hypotéza zamítnuta.
 - i. p – Left – odkazuje na poměr velikosti efektů, které jsou menší než vypočtená velikost efektu daného pozorování.
 - ii. p – Right – odkazuje na poměr velikosti efektů, které jsou větší než vypočtená velikost efektu daného pozorování.
 - iii. p – Tot – označuje oboustrannou hodnotu p, odkazuje na poměr velikostí efektů absolutní hodnoty, které jsou extrémnější než vypočtená velikost efektu daného pozorování.

g. SE = standard error

- SE je pojem označující v angličtině standard error (směrodatná odchylka), jedná se o nepřesnost měření při výběrovém šetření, objevuje se v případě, pokud je počítáno pouze se vzorkem z několika hodnot (v případě bakalářské práce s průměrem hodnot).

h. Cohesion categories

- Koheze je určitou silou, jež drží v tomto případě kategorie pohromadě, čím větší tato hodnota je, tím více jsou si hodnoty kategorie podobné.

3.6 Analýzy v programu R

Pro vyhotovení interpreovatelných výsledků je použit program R (RStudio). Stejně jako všechny předchozí kódy, tak i tento je převzatý od původních tvůrců. Konkrétněji jsou hlavně

použity určité řádky původního dokumentu R (678–1026), jelikož dokument obsahuje kód analýzy pro všechny 3 studie celého publikovaného výzkumu. R skript je pouze trochu upraven, jelikož jsem našla pár nesrovnalostí, které na mém počítači nefungovaly a díky nim analýza nemohla proběhnout.

3.6.1 Nesrovnalosti a nutné úpravy

Tabulka 1: Nesrovnalosti a nutné úpravy

nesrovnalost	následek	úprava	řádek
Přebytečný #	Anulace kódu	Odstranění #	28, 424, 680, 949, 2042, 2786, 3121, 3296, 3464, 3500, 3676, 3717, 3739
Load(data)	Moje R to nezná, nenačte se soubor s daty	Místo load dát source	23
Formula(x) = formula(time)	Analýza nepokračuje, regrese pro čas neproběhne	Místo formula(time) dát formula(paste(time))	382, 631
Absence kódu pro zvýšení overlaps	Program R nepokračuje a napíše: „too many overlaps“ a nějaká data by mohla být vyloučena	Přidání kódu „options(ggrepel.max.overlaps = Inf)“	Kdekoliv
Špatně zvolené pořadí barev v kódu určeném pro tvorbu grafu	Ve výsledném grafu nepasují barvy označující intervaly v legendě s barvami přímk v grafu	Barva „seagreen“ se musí vyměnit s barvou „red“	942

Zdroj: vlastní zpracování

Zmíněny jsou řádky původního dokumentu, a jelikož jsem přidávala nějaké kódy pro dodatečné analýzy a dovysvětlení grafů, je možné, že se tyto hodnoty budou lišit od skriptu R Studia, jenž je přiložen.

3.6.2 Hlavní korpusy pro výzkum a povolání v nich

Do programu R se načte excelový soubor čítající data ze všech korpusů vygenerovaný v předchozí části. Stanoví se hlavní korpusy na základě nejvyšší koncentrace zmínek o povolání a zanalyzuje se, jaká povolání jsou nejčastější, jaká vykazují nejsilnější efekty a zároveň se pro tyto hodnoty nalezne průměr.

3.6.3 Regresní analýza

Regrese se použije ve chvíli, kdy se zjišťují rozdíly efektů v čase získání dat a rozdíly efektů díky odlišné věkové kategorii.

3.6.4 Binomické testy

Jeden binomický test se provede pro zjištění, které atributy (povolání) jsou signifikantní pro $p = 0,05$. Zjistí se tedy počet profesí, které mezi sebou mají systematický vztah na hladině významnosti 0,05.

Za pomoci binomických testů se také ověří počet jednotlivých povolání v daných intervalech WEATU a také zkontroluje správnost rozdělení podílu mužských a ženských povolání v celkovém výčtu 82 povolání.

3.6.5 Korelace

Korelace je použita při zjištění vztahu mezi reálnými statistikami a výsledky analýz bakalářské práce. Zkoumá vztah mezi četností genderových stereotypů o povolání v textu a reálném podílu jednotlivého genderu v zaměstnanosti jednotlivých profesí. Korelace se provádí, jak pro průměrné výsledky, tak pro výsledky jednotlivých korpusů.

3.6.6 Kovariance

Kovariance je použita v případě ukázání souvislosti a lineárního vztahu mezi věkovými kategoriemi a hodnotami WEATu.

4 Výsledky

Výsledky jsou ukázány na základě analýz v programu R (RStudio), kde hlavními metodami jsou regrese, binomické testy, korelace a kovariance.

4.1 Celková data

Celková data předpokládá 12 korpusu, a jelikož je to docela vysoké číslo, zvolilo se 7 hlavních, které obsahují nejvyšší koncentraci zmínek o povolání. S těmito 7 hlavními korpusy se bude pracovat v následujících analýzách.

4.1.1 Výskyt povolání v korpusech

Hned na začátku je zjištěno, že v alespoň 1 ze všech 12 korpusů se vyskytují zmínky o dohromady 84 profesích. Následně analýza ukázala, že nejvyšší koncentraci pracovních pozic obsahuje 7 korpusů a jsou jimi Mluva dětí, Mluva mířená dětem, Dětské knihy, TV dětí – kombinace, Mluva dospělých, Gutenberg a TV dospělých Simply scripts. V každém z nich se aspoň jednou vyskytuje 17 nejčastějších profesí, které se opakují. Pokud ovšem snížíme počet výskytů a budeme předpokládat obsah např. v 5 korpusech, tak v nich se opakuje profesí 39, ve 4 korpusech je to pak 50 pozic a v aspoň 3 korpusech 60 pozic. Je tedy ukázáno, že ne každý korpus je, co se týče obsahu zkoumaných profesí, stejný a dochází k menším rozdílům.

Pro lepší demonstraci výskytu povolání v jednotlivých korpusech je uveden případ, kdy se předpokládá výskyt profese v alespoň 5 korpusech.

Tabulka 2: Rozdělení povolání, které se vyskytly alespoň v 5 korpusech, mezi jednotlivými korpusy

korpus	Mluva dospělých	Gutenberg	Dětské knihy	Dětská mluva	Mluva mířená dětem	TV dětí – kombinace	TV dospělých SimplyScripts
počet výskytů	37	39	30	26	34	39	34

Zdroj: vlastní zpracování

Příčemž v tabulce jsou ukázány dané korpusy a kolik se v nich aktuálně po omezení vyskytuje pracovních pozic. Nejvíce koncentrovanými korpusy jsou Gutenberg a TV dětí – kombinace s 39 výskytů a nejméně korpus Dětské mluvy.

Následující tabulka ukazuje výskyt jednotlivých profesí ve všech 7 korpusech. Dohromady je napočítáno 344 výskytů.

Tabulka 3: Četnost 82 zkoumaných povolání v 7 hlavních korpusech

accountant	actor	adjuster	administrator	agriculturalist	analyst	announcer	architect
4	7	1	4	1	3	4	5
assembler	athlete	author	baker	barber	bartender	biologist	bookkeeper
2	5	6	7	5	3	2	1
butcher	cabinetmaker	carpenter	cleaner	clergy	clerks	cook	counselor
7	1	6	7	1	1	7	4
dentist	designer	dispatcher	dressmaker	economist	editor	electrician	engineer
6	4	2	2	1	5	3	7
excavator	executive	farmer	firefighter	fisher	gardener	geologist	glazier
3	4	7	3	6	6	1	1
grader	guard	hairstylist	chemist	inspector	interviewer	janitor	laborer
5	7	2	2	4	3	5	2
lawyer	librarian	lumberjack	machinist	maid	manager	mechanic	messenger
5	6	2	1	7	7	5	5
model	musician	nurse	painter	pharmacist	photographer	physician	pilot
7	6	7	6	2	5	5	7
plumber	police	professional	programmer	psychologist	receptionist	sheriff	supervisor
6	7	6	1	3	3	7	4
tailor	teacher	technician	therapist	underwriter	upholsterer	veterinarian	waitress
4	7	4	3	1	1	4	5
welder	writer						
3	7						

Zdroj: vlastní zpracování

Přičemž jejich celkové rozdělení v 7 hlavních korpusech vypadá takto.

Tabulka 4: Celkový počet zmínek o povolání v korpusech – součet pro korpusy

korpus	Mluva dospělých	Gutenberg	Dětské knihy	Dětská mluva	Mluva mířená dětem	TV dětí – kombinace	TV dospělých SimplyScripts
počet výskytů	51	76	32	28	39	61	57

Zdroj: vlastní zpracování

Lze pozorovat, že největší výskyt zmínek o povolání je v korpusu Gutenberg a TV dětí – kombinace, což je logické, jelikož jsou to korpusy největší. Naopak nejmenší výskyt profesí je v korpusu Dětské mluvy. Data pro tento korpus byla sesbírána u dětí ve věku 0–12 let a

průměrný věk byl 2,92 let, což napovídá, že děti ve svých konverzacích nenarážely na povolání tak moc jako například dospělí.

4.1.2 Efekty povolání v korpusech

Průměrný efekt z celého souboru napříč korpusem je 0,03. Jelikož se číslo silně blíží 0, lze upozorovat na vyváženost stereotypů mužů i žen. Ovšem to, že je číslo kladné, ukazuje na vyšší výskyt stereotypů týkajících se mužů. Hodnoty jednotlivých efektů se pohybují v rozpětí -1.51 až 1.34, a jak už bylo zmíněno, kladná hodnota přísluší mužům a záporná ženám. Zároveň čím vyšší hodnota efektu je, tím se jedná o silnější stereotyp. Nejsilnější efekt, nejsilnější genderový stereotyp o povolání je tedy spojován s ženami. Tyto krajní hodnoty nejsou nejpočetnější.

Každému jednotlivému povolání je také spočítán průměrný efekt asociace s povoláním, který je zobrazen v další tabulce. Jedná se o součet jednotlivých efektů povolání z každého korpusu vydělen počtem výskytu povolání.

Tabulka 5: Průměrné efekty povolání

accountant	actor	adjuster	administrator	agriculturalist	analyst	announcer	architect
-0,07	0,29	-0,48	0,45	0,75	0,45	-0,06	0,43
assembler	athlete	author	baker	barber	bartender	biologist	bookkeeper
0,49	0,46	-0,07	-0,45	0,24	-0,19	0,49	0,24
butcher	cabinetmaker	carpenter	cleaner	clergy	clerks	cook	counselor
-0,15	0,04	0,11	-0,04	0,21	0,12	-0,22	-0,26
dentist	designer	dispatcher	dressmaker	economist	editor	electrician	engineer
-0,16	0,28	-0,03	-0,87	0,51	0,12	0,36	0,63
excavator	executive	farmer	firefighter	fisher	gardener	geologist	glazier
0,80	0,12	-0,08	0,13	0,13	0,02	0,70	0,21
grader	guard	hairdresser	chemist	inspector	interviewer	janitor	laborer
0,18	0,61	-0,10	0,42	0,62	-0,41	0,12	0,31
lawyer	librarian	lumberjack	machinist	maid	manager	mechanic	messenger
0,10	-0,54	0,25	0,95	-0,86	0,14	0,47	0,25
model	musician	nurse	painter	pharmacist	photographer	physician	pilot
-0,17	0,13	-0,99	-0,16	0,58	-0,22	-0,06	0,71
plumber	police	professional	programmer	psychologist	receptionist	sheriff	supervisor
0,41	0,18	0,12	-0,45	0,09	-0,09	-0,07	0,41
tailor	technician	technician	therapist	underwriter	upholsterer	veterinarian	waitress
0,51	-0,74	0,17	-0,24	0,01	-0,38	-0,58	-0,56
welder	writer						
0,04	-0,26						

Zdroj: vlastní zpracování

Na následujícím grafu můžeme vidět seřazené profese podle průměrné velikosti efektů od nejvíce mužských po nejvíce ženské. Zároveň je na něm vyznačen rozptyl velikosti efektů, kterého jednotlivé povolání dosáhlo v různých korpusech.

Z výsledků lze taky vyčíst, v jakých korpusech převládají mužské a v jakých ženské stereotypy, zde je zastoupení průměrných efektů v korpusech.

Tabulka 6: Průměrné efekty povolání v korpusech

korpus	Mluva dospělých	Gutenberg	Dětské knihy	Dětská mluva	Mluva mířená dětem	TV dětí – kombinace	TV dospělých SimplyScripts
průměrný efekt	-0.14	0.16	0.32	-0.01	-0.03	0.07	-0.10

Zdroj: vlastní zpracování

Nejvyšší kladný průměrný efekt se vyskytuje u korpusu Dětské knihy s hodnotou 0.32. Možným vysvětlením může být to, že v dětských pohádkách se mezi hlavními postavami častěji vyskytují právě muži praktikující určitá povolání. I když se jedná o anglickou literaturu, v České republice bychom mohli zmínit pohádku O statečném kováři nebo O létajícím ševci. Naproti tomu nejvyšší záporný efekt se ukazuje u mluvy dospělých, a jelikož je to korpus vytvořený na základě telefonních rozhovorů, je možné, že v těchto dialogích dominovaly ženy, které se považují za výřečnější a jejich témata mohly častěji být informace ohledně dalších žen a zjevně i o povoláních.

4.2 Popis meta-analýzy

Meta-analýzou rozumíme sadu metod, které se používají v případě analýz dat z různých studií, zdrojů. V případě bakalářské práce se jedná o kombinaci dat z různých korpusů. Díky meta-analýze lze dosáhnout vyšší statistické síly a je také užitečná pro určení a případné prozkoumání zdrojů zkreslení a kvantifikaci heterogenity mezi zdroji dat a studiemi (Tatsioni, & Ioannidis, 2017, s. 117).

4.2.1 Nejsilněji asociovaná povolání s genderem napříč korpusy

Průměrné hodnoty efektů povolání (vypočteny ze všech 7 hlavních korpusů) se nacházejí v intervalu od -0.99 do 0.95. Přičemž, jak už bylo zmíněno, kladné hodnoty přísluší efektům týkajících se povolání, která jsou vnímána jako mužská a záporné hodnoty zase přísluší profesím spíše ženského typu.

Pokud se podíváme na šest nejsilnějších mužských povolání, dostaneme názvy: machinist (opravář / údržbář strojů), excavator (bagrista), agriculturalist (agrotechnik či zemědělský odborník), geologist (geolog), pilot (pilot) a inspector (inspektor).

Při zaměření se na šest „nejvíce ženských profesí“, dostaneme výsledky: nurse (zdravotní sestra), maid (služebná či pokojská), dressmaker (švadlena), teacher (učitelka), veterinarian (veterinářka) a waitress (servírka).

4.2.2 Zastoupení profesí v jednotlivých intervalech velikosti efektu

Za pomoci binomického testu s původně nastavenou hladinou významnosti na 5 % je zjištěno, že pouze 22 povolání je signifikantních. Zbylých 60 je nesignifikantních, což značí, že pouze 26,83 % proměnných (povolání) je v binomickém testu úspěšných a lze mezi nimi prokázat systematický vztah.

Tabulka 7: Zastoupení signifikantních povolání

	Zastoupení profesí v rámci celku	Zastoupení profesí v rámci intervalu
Muži	14 (17,07 %)	14 (63,64 %)
Ženy	8 (9,76 %)	8 (36,36 %)
Celkem	22 (26,83 %)	22 (100 %)

Zdroj: vlastní zpracování

Po zaměření se na intervaly efektů jednotlivých povolání zjistíme, že v intervalu $[-0,1; 0,1]$ se vyskytuje 20,73 % povolání z celkových 82 (v přepočtu na čísla - 17 profesí). Jedná se o pracovní pozice, které jsou z celkových 82 profesí nejslaběji spojovány s mužským či ženským pohlavím a nejsou tedy vysoce signifikantní. Povolání, která spadají mimo interval hodnoty efektu $[-0,1; 0,1]$, je 79,27 % z celkových 82 (65 profesí) a ty už jsou s pohlavím spojována silněji.

Tabulka 8: Zastoupení povolání mimo interval hodnot efektů $[-0,1; 0,1]$

	Zastoupení profesí v rámci celku	Zastoupení profesí v rámci intervalu
Muži	42 (51,22 %)	42 (64,62 %)
Ženy	23 (28,05 %)	23 (35,38 %)
Celkem	65 (79,27 %)	65 (100 %)

Zdroj: vlastní zpracování

Pokud zpřísníme kritérium a podíváme se mimo interval hodnoty efektu $[-0,2; 0,2]$, dostaneme 47 povolání z 82 (57,32 %), která jsou úspěšná a splňují podmínku. A rozdělení mezi muži a ženami vypadá takto.

Tabulka 9: Zastoupení povolání mimo interval hodnoty efektů [-0,2; 0,2]

	Zastoupení profesí v rámci celku	Zastoupení profesí v rámci intervalu
Muži	30 (36,59 %)	30 (63,83 %)
Ženy	17 (20,73 %)	17 (36,17 %)
Celkem	47 (57,32 %)	47 (100 %)

Zdroj: vlastní zpracování

A když se zaměříme na povolání mimo interval hodnoty efektu [-0,3; 0,3], dostaneme pouze 36 povolání z 82 (43,90 %), což ukazuje, že jednotlivé efekty z větší části spadají do intervalu [-0,3; 0,3].

Tabulka 10: Zastoupení povolání mimo interval hodnoty efektů [-0,3; 0,3]

	Zastoupení profesí v rámci celku	Zastoupení profesí v rámci intervalu
Muži	24 (29,27 %)	24 (66,67 %)
Ženy	12 (14,63 %)	12 (33,33 %)
Celkem	36 (43,90 %)	36 (100 %)

Zdroj: vlastní zpracování

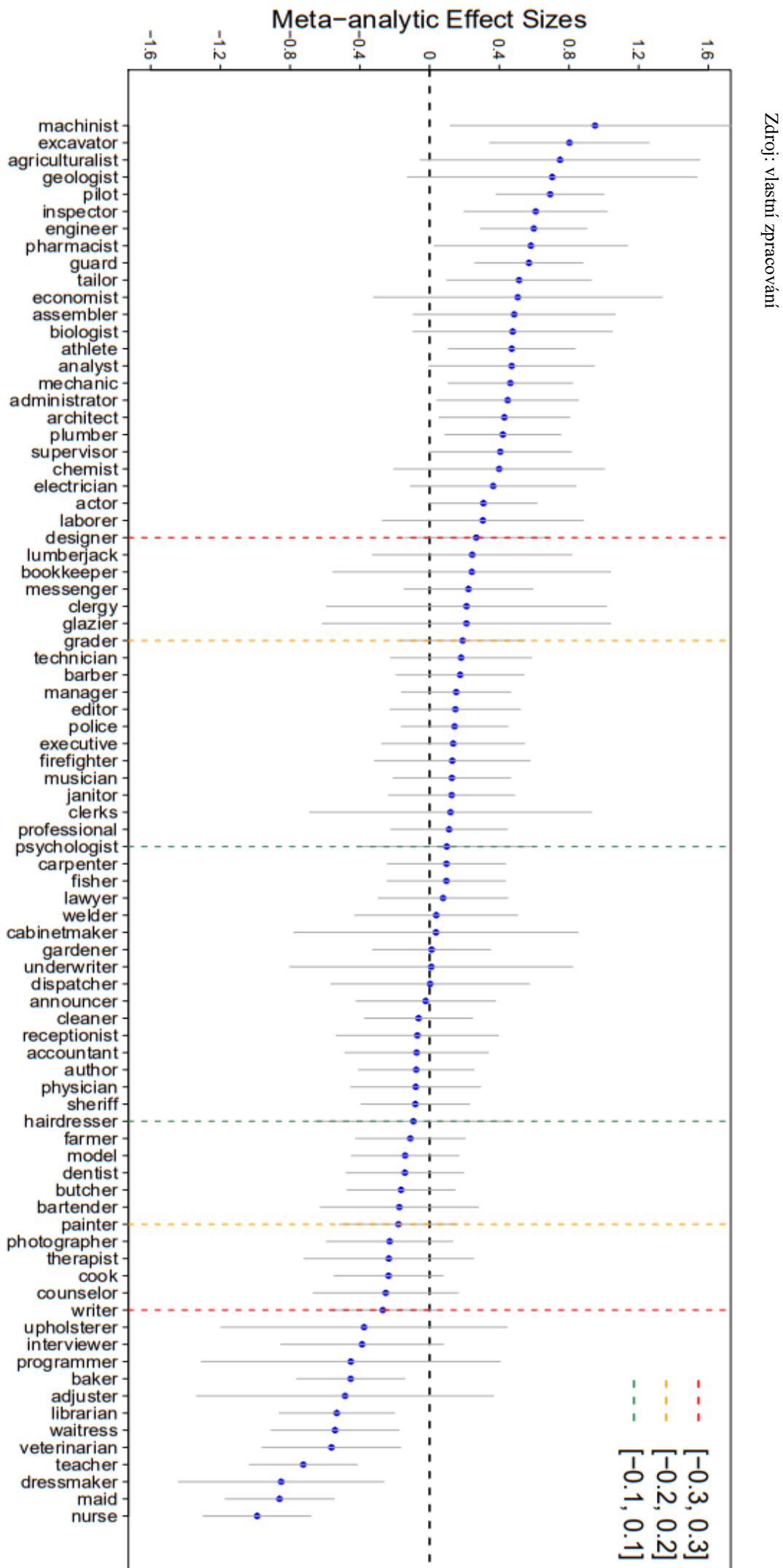
V tomto posledním příkladě binomického testu je demonstrováno, že nejvýznamnější a nejsilnější efekty povolání přísluší mužskému pohlaví. Toto zjištění odkazuje na další stereotyp, a to, že muži jsou častěji asociováni s prací.

Pokud se podíváme na všech 82 profesí, tak za pomoci dalšího binomického testu předchozí zjištění potvrdíme. Výsledek čítá 51 mužských profesí, což je 62 % z celku (na hladině významnosti $p = 0,04$). Tento výsledek je shodný se statistikami U.S. Bureau of Labor Statistics (1998), která ve svém výzkumu z roku 1999 (rok byl vybrán díky podobnosti dostupnosti zkoumaných dat), který ukazuje, že pracovní síla z 60 % byla tvořena muži.

Zhodnocení

Po zhodnocení výsledků jednotlivých binomických testů je patrné, že největších efektů, a tudíž i nejsilnějších asociací povolání s pohlavím dosahují muži. Při postupném zpřísňování kritérií a nastavování vyšších absolutních hodnot intervalu efektů se i jejich zastoupení v rámci intervalu lehce zvyšuje.

Na následujícím grafu dále můžeme spatřit jednotlivá povolání hodnocena podle průměrného efektu napříč 7 hlavními korpusy. Jsou v něm vyznačeny jak průměrné efekty, tak i rozptyl nejvyšších a nejnižších hodnot efektů pro dané povolání. Opět platí, že kladné hodnoty efektů přísluší mužům a záporné ženám. Zároveň jsou v grafu vyznačené intervaly efektů [-1; 1], [-2; 2] a [-3; 3] pro lepší přehlednost hodnoty průměrného efektu profese a jejího umístění na škále hodnot v grafu.



4.2.3 Nejsilnější genderové stereotypy o povolání v jednotlivých korpusech

Jak už bylo v předchozích tabulkách ukázáno, v každé kategorii (v každém korpusu) se vyskytuje jiná koncentrace zmínek o povolání, a i trochu jiné síly efektů. To znamená, že v jednotlivých korpusech mohou mít jednotlivá povolání jiný průměrný efekt než v celku 7 korpusů. V následujících tabulkách je sestupně zobrazeno prvních 6 povolání s nejsilnějšími efekty pro jednotlivé hlavní korpusy a pohlaví:

1. Mluva dětí

Tabulka 11: Povolání s nejsilnějšími efekty pro korpus Mluva dětí

Muži	manager	pilot	excavator	plumber	grader	guard
Efekty mužů	0,92	0,80	0,78	0,77	0,69	0,63
Ženy	maid	nurse	teacher	sheriff	cleaner	cook
Efekty žen	-1,33	-1,18	-0,77	-0,59	-0,49	-0,44

Zdroj: vlastní zpracování

2. Mluva mířená dětem

Tabulka 12: Povolání s nejsilnějšími efekty pro korpus Mluva mířená dětem

Muži	athlete	plumber	gardener	excavator	firefighter	announcer
Efekty mužů	1,07	0,89	0,89	0,75	0,56	0,55
Ženy	nurse	librarian	cook	maid	veterinarian	teacher
Efekty žen	-1,30	-0,91	-0,87	-0,80	-0,71	-0,69

Zdroj: vlastní zpracování

3. Dětské knihy

Tabulka 13: Povolání s nejsilnějšími efekty pro korpus Dětské knihy

Muži	police	engineer	barber	guard	musician	pilot
Efekty mužů	1,34	1,34	1,32	1,24	1,18	1,15
Ženy	dressmaker	maid	nurse	teacher	model	baker
Efekty žen	-1,08	-0,97	-0,93	-0,91	-0,62	-0,23

Zdroj: vlastní zpracování

4. TV dětí – kombinace

Tabulka 14: Povolání s nejsilnějšími efekty pro korpus TV dětí – kombinace

Muži	actor	cook	pilot	janitor	inspector	architect
Efekty mužů	1,15	1,07	1,01	0,79	0,61	0,57
Ženy	veterinarian	librarian	nurse	farmer	baker	gardener
Efekty žen	-1,01	-0,90	-0,84	-0,76	-0,71	-0,55

Zdroj: vlastní zpracování

5. Mluva dospělých

Tabulka 15: Povolání s nejsilnějšími efekty pro korpus Mluva dospělých

Muži	mechanic	guard	inspector	athlete	chemist	pilot
Efekty mužů	0,99	0,97	0,94	0,88	0,83	0,78
Ženy	librarian	nurse	waitress	teacher	musician	editor
Efekty žen	-1,51	-1,21	-1,18	-1,10	-1,06	-1,03

Zdroj: vlastní zpracování

6. Knihy pro dospělé

Tabulka 16: Povolání s nejsilnějšími efekty pro korpus Knihy pro dospělé

Muži	machinist	administrator	excavator	editor	engineer	agriculturalist
Efekty mužů	0,95	0,94	0,88	0,87	0,78	0,75
Ženy	maid	waitress	nurse	dressmaker	adjuster	upholsterer
Efekty žen	-1,24	-1,10	-0,84	-0,65	-0,48	-0,38

Zdroj: vlastní zpracování

7. TV dospělých Simply Scripts

Tabulka 17: Povolání s nejsilnějšími efekty pro korpus TV dospělých Simply Scripts

Muži	analyst	guard	pilot	messenger	assembler	tailor
Efekty mužů	1,02	0,75	0,72	0,60	0,52	0,50
Ženy	interviewer	cook	teacher	model	maid	announcer
Efekty žen	-0,93	-0,90	-0,88	-0,83	-0,83	-0,81

Zdroj: vlastní zpracování

A. Mužská povolání

- Z výsledků lze vyčíst, že mezi nejsilnější zmiňované povolání asociované s mužským pohlavím patří pilot (pilot), který se umístil mezi nejsilnějšími 6 profesemi v 5 ze 7 korpusů. Dále vidíme povolání guard (strážný, dozorce), který se vyskytuje ve 4 ze 7 korpusů a dalším častým je excavator (bagrista), kterého najdeme ve 3 ze 7 korpusů.
- Ve výsledcích původní práce je zmíněno, že profese guard se vyskytuje z hlediska efektů mezi 6 nejsilnějšími profesemi ve 3 ze 7 korpusů, v mé analýze se objevila ve 4 ze 7 korpusů, a jelikož původní práci replikují a všechny analýzy mi vycházejí stejně a číselné hodnoty jsou prakticky identické, je tato malá nesourodost nutná zmínit.

B. Ženská povolání

- Mezi povolání s nejsilnějšími efekty spojovanými s ženami patří nurse (zdravotní sestra), která se vyskytuje mezi 6 nejsilnější asociovanými profesemi s ženským pohlavím v 6 ze 7 korpusů. Dále maid (služebná či pokojská) a

teacher (učitelka) se umístily mezi 6 nejsilnějšími profesemi spojovanými s ženami v 5 ze 7 korpusů. Dalším silně asociovaným pohlavím s ženami je librarian (knihovnice), která se na prvních příčkách objevila ve 3 ze 7 korpusů.

C. Povolání spojená s oběma pohlavími

- Povolání cook (kuchař / kuchařka) se silněji asociovalo s ženským pohlavím, kde se vyskytovalo v 3 ze 7 korpusů, ovšem objevilo se i v 1 korpusu (Dětská TV) u mužů. Další podobně obojetnou profesí je gardener (zahradník / zahradnice), která byla s muži silně asociována v případě 1 korpusu (Mluva mířená dětem) a s ženami také v 1 korpusu (Dětská TV), její průměrný efekt ovšem naznačuje vyšší spojitost s muži nebo také povolání musician (hudebník / hudebnice).

4.2.4 Meta-regrese napříč korpusy, rozdíly ve věkové kategorii

Velikost efektů pro genderové stereotypy z hlediska povolání byla porovnána i pro věkové kategorie. Všech 7 korpusů je možno rozdělit na dvě skupiny – buď zkoumají přirozený jazyk dětí (či určený dětem) nebo dospělých. Do dětských korpusů patří Mluva dětí, Mluva mířená dětem, Dětské knihy, TV dětí – kombinace a korpusy o dospělých jsou Mluva dospělých, Gutenberg a TV dospělých Simply Scripts.

Regrese ukázala, že rozdíly jsou nesignifikantní ($b = -0,08$; 95% CI = $[-0,19; 0,04]$; $z = -1,27$; $p = 0,20$). Hladina významnosti p je nad 0,05, proto se nulová hypotéza nezamítá a výsledky jsou nesignifikantní.

Rozdíly ve velikostech efektů jsou tedy v obou skupinách, jak pro děti, tak dospělé nepatrné. Přirozený jazyk není nijak přizpůsobován pro určitý věk, popisuje realitu, která je stejná pro obě věkové skupiny.

4.2.5 Meta-regrese napříč korpusy, rozdíly v čase získání dat

Všech 7 hlavních korpusů se dá rozdělit i na základě doby pořízení dat, některé jsou starší, jiné novější a některé se řadí doprostřed. Mezi starší korpusy patří Dětské knihy a Gutenberg, mezi střední korpusy zařadíme Mluvu dětí, Mluvu mířenou dětem a Mluvu dospělých a mezi korpusy, které mapují novější data se řadí TV dětí – kombinace a TV dospělých Simply Scripts.

Regresní analýza zjistila, že starší korpusy zobrazují více genderových stereotypů o povolání spojených s muži a postupně se tato síla efektů snižuje. Pro starší korpusy vyšla regrese takto ($b = 0,20$; 95% CI = $[0,10; 0,31]$; $z = 3,84$; $p < 0,0001$), pro střední ($b = -0,27$; 95% CI = $[-0,41; -0,13]$; $z = -3,75$; $p = 0,0002$) a pro nejmladší a ($b = -0,21$; 95% CI = $[-0,35; -0,07]$; $z = -2,98$; $p = 0,003$).

Je patrné, že p – hodnota se postupně zvyšuje a s tím i genderová vyrovnanost, co se týče výskytu stereotypů o povolání. Z výsledků lze odvodit, že ve starších korpusech se muži více asociovali s prací a postupně s emancipací žen a větším zařazením žen do pracovního života se tato skutečnost odráží i v přirozeném jazyce.

4.2.6 Korelace se skutečnými statistikami (U.S. Bureau of Labor Statistics)

Bakalářská práce a původní studie se zároveň opírají o statistiky skutečné pracovní síly, které jsou dostupné na webových stránkách americké statistické společnosti U.S. Bureau of Labor Statistics (1998). Konkrétně o dokument s 307 povoláními (40 % profesí je ženských a 60 % mužských), zároveň je v něm zobrazeno, kolik procent v dané profesi zabírají ženy.

Byla prokázána silná pozitivní korelace mezi genderovými stereotypy o povolání a reálnými statistikami BLS ($r = 0,53$; 95% CI = $[0,35; 0,67]$; $t(80) = 5,59$; $p < 0,001$). Tato korelace je konkrétně vyhotovena mezi průměrným efektem všech povolání ve všech korpusech a zastoupením mužů v reálných statistikách. Korelace tedy prokázala, že čím vyšší je procento zastoupení mužů v statistikách BLS, tím silnější asociace existuje mezi slovem muž (či slovem zástupným) a nějakou profesí, toto zjištění se dá vztáhnout i na pohlaví ženské. Jedná se o lineární závislost.

Dále jsou vypočítány korelace pro průměrné efekty povolání v jednotlivých korpusech a podíly mužů v reálné zaměstnanosti určitých profesí v statistikách BLS. Korelace se nacházejí v intervalu $[0,21; 0,78]$:

Tabulka 18: Korelace mezi průměrnými efekty povolání a procentuálním zastoupením genderu v reálné zaměstnanosti povolání mezi korpusem

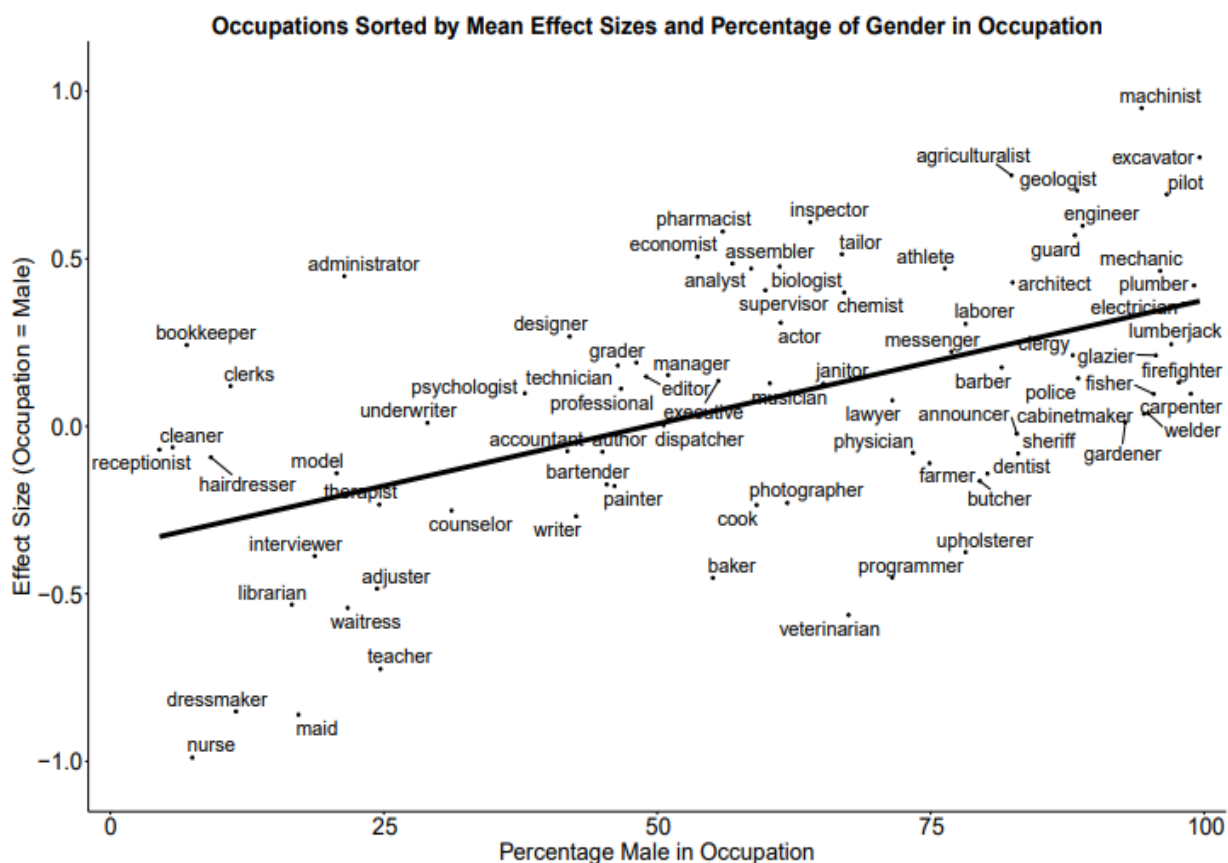
korpus	Mluva dospělých	Gutenberg	Dětské knihy	Dětská mluva	Mluva mířená dětem	TV dětí – kombinace	TV dospělých SimplyScripts
korelace	0,43	0,39	0,78	0,46	0,59	0,21	0,45
p	0,0017	0,0006	<0,0001	0,0136	0,0001	0,1020	0,0004

Zdroj: vlastní zpracování

Nejsilnější korelace byla prokázána v korpusu Dětských knih s hodnotou 0,78 a nejnižší zase u TV dětí – kombinace s hodnotou 0,21. V obou případech se jedná o dětské korpuse.

Všechny korelace jsou kladné, jedná se tedy o přímou závislost (pozor, ne kauzalitu!). Pokud se tedy hodnota jedné proměnné zvýší, zvýší se i hodnota druhé. Kdyby korelace byly záporné, tak po zvýšení hodnoty jedné veličiny by se snížila hodnota druhé.

Graf 1: Povolání rozdělená dle efektu a procentuálního zastoupení genderu v reálné zaměstnanosti povolání



Zdroj: vlastní zpracování

Na grafu vidíme prostor, ve kterém jsou umístěna zkoumaná povolání. Na ose x se nachází procentuální zastoupení mužů v jednotlivých profesích z reálných statistika a na ose y jsou hodnoty průměrných efektů profesí, přičemž kladné hodnoty vyjadřují vyšší asociace povolání s muži a záporné s ženami.

V kapitole Nejsilněji asociovaná povolání s genderem napříč korpusy jsou ukázány nejvíce spojovaná povolání s pohlavími na základě průměrného efektu ze všech korpusů, což se odráží i v grafu. Můžeme vidět, že nejvíce mužné povolání – machinist (opravář / údržbář strojů; efekt 0,95), se nachází vpravo nejvíce nahoře, nicméně podle statistik BLS je nejmužnější excavator (bagrista), a proto se nachází v grafu více vpravo, blíže k hodnotě 100 na ose x. Excavator (bagrista) vykazoval efekt 0,80, což také dokazuje vysokou asociaci s mužským pohlavím, a proto jsou tato dvě povolání tak blízko sebe a opět lze říct, že reálné statistiky o zastoupení pohlaví v povoláních se odrážejí v genderových stereotypech v textu přirozeného jazyka. V případě ženského pohlaví se statistiky i výsledky bakalářské práce shodly, že nejžensetější povolání je nurse (zdravotní sestra), nicméně na druhé a třetí příčce došlo k obměně. Dle statistik je druhá nejvíce asociovaná profese s ženami dressmaker (švadlena) a třetí maid (pokojská) a v bakalářské práci je to na základě efektů naopak, nicméně rozdíl není tak veliký, proto se může potvrdit silná podobnost výsledků s realitou.

4.3 Vyhodnocení hypotéz

V následující kapitole se zaměřím na jednotlivé hypotézy, které byly stanoveny v teoretické části a na základě výsledků je zhodnotím.

H1: Existují konzistentní rozdíly ve vnímání genderových stereotypů o povolání u mužů i žen v jednotlivých oblastech (napříč korpusy).

Správnost hypotézy je ověřena v kapitole Nejsilnější genderové stereotypy o povolání v jednotlivých korpusech, kde můžeme vidět, že v jednotlivých korpusech se na místě nejsilněji spojovaných povolání s daným genderem objevují často ty samé profese. Jejich opakování potvrzuje tvrzení, že existují konzistentní rozdíly napříč korpusy.

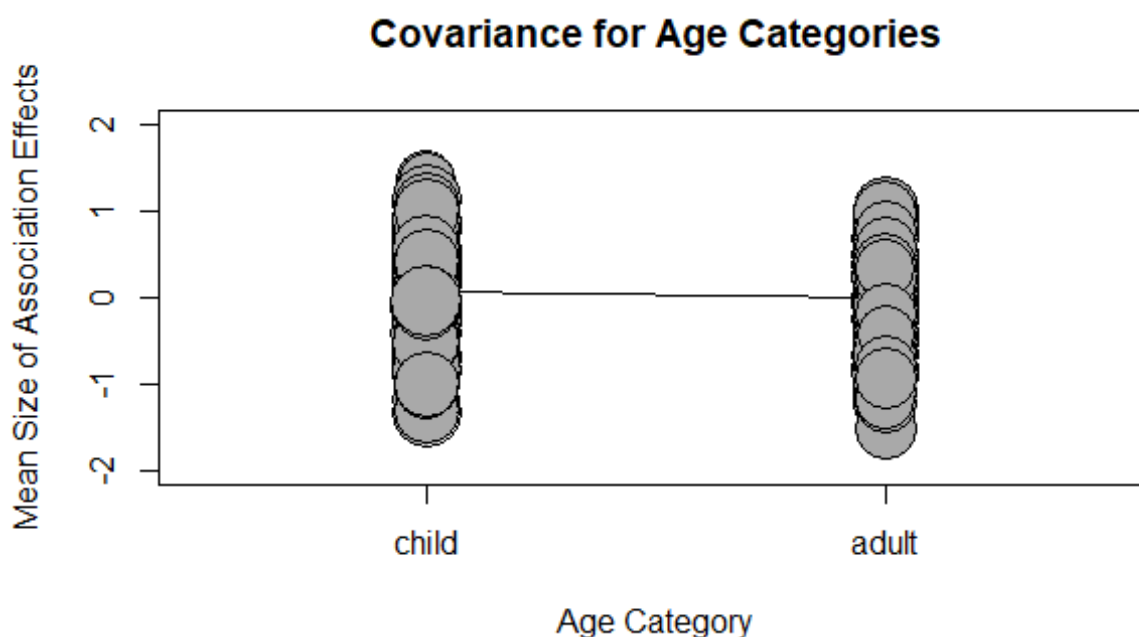
Nicméně je taky pravda, že ne v každém korpusu je daná profese vnímána stejně, o čemž napovídá efekt povolání, který se s daným korpusem občas liší. Někdy se na prvních příčkách v určitém korpusu umístí povolání, které v ostatních korpusech umístěno není. Tento fakt řeší průměr efektů jednotlivých povolání, kde je ve finále zjištěno, zda profese je spíše spojována s mužem či ženou. A aby se úplně předešlo tomuto případnému zkreslení, bylo by potřeba získat ještě objemnější datasety, které by vše sjednotili. Každopádně pro výsledky studie a detekování genderových stereotypů o povolání v přirozeném jazyce jsou zkoumaná data postačující.

H2: V komunikaci rozdělené podle věkové kategorie se vyskytují podobně silné genderové stereotypy o povolání a také v podobném zastoupení.

Na tuto otázku – hypotézu odpovídá kapitola Meta-regrese napříč korpusy, rozdíly ve věkové kategorii a bylo zjištěno, že prakticky neexistují rozdíly v zastoupení stereotypů ani v průměrné síle asociací stereotypů o povolání s jednotlivým pohlavím v případě věkových kategorií. Hlavním indikátorem tohoto zjištění je p – hodnota, která dosáhla úrovně 0,20, a proto byla tato hypotéza potvrzena.

Pro další potvrzení zjištění je přiložen bublinový graf z programu R, který ukazuje míru kovariance pro věkové kategorie. A jelikož kovariance počítá s odhady středních hodnot veličin, tak z grafu můžeme vyčíst, že střední hodnoty efektů asociací mezi povoláním a pohlavím jsou, jak v případě dětské, tak i dospělé věkové kategorie opravdu podobné stejně jako jejich zastoupení, jež je zobrazeno v podobě bublin.

Graf 2: Kovariance na základě WEATu pro věkové kategorie



Zdroj: vlastní zpracování

H3: Genderové stereotypy o povolání v přirozeném jazyce korelují s administrativními reálnými statistikami o zaměstnanosti mužů a žen v jednotlivých povoláních.

Pro potvrzení této hypotézy bylo zapotřebí použít data, která pojednávají o rozdělení reálné zaměstnanosti a rozdělení mezi pohlavími. Proto jsou použity statistiky společnosti BLS. V kapitole Korelace se skutečnými statistikami (U.S. Bureau of Labor Statistics) se čtenář dozví, že korelace dosáhla úrovně 0,53, což značí o silné pozitivní korelaci a tím pádem i silného vztahu mezi dvěma zmíněnými veličinami. V datasetu U.S. Bureau of Labor Statistics (1998) je celkem 307 povolání a 60 % z nich častěji vykonávají muži, 40 % poté připadá ženám.

Ve výsledcích práce se dozvíme, že v textech přirozeného jazyka je 62 % povolání častěji spojována s muži a 38 % s ženami. Toto jsou ty genderové stereotypy o povolání, to, co je implicitně zakotveno v textu.

A i když se jedná o genderové stereotypy spojené s povoláními, je zároveň patrné, že se opírají o skutečnost. Skutečnost reality se odráží v mluvě, knihách a audiovizuálních médiích.

H4: Na základě dostupných dat zkoumaného článku lze danou studii replikovat a dojít ke stejným výsledkům.

Jelikož je replikovaný článek velmi dobře popsán a veškerá data jsou pro něj zpřístupněna na serveru OSF, tak tato hypotéza byla potvrzena. Nicméně, jak už bylo zmíněno v kapitole Child - directed audiovisual media = TV dětí, data od společnosti Disney se nenalezla, proto musely být využity již zpracované word vectory pro tento korpus. Vše ostatní k dispozici bylo a výzkum mohl proběhnout přesně podle stejných kroků jako původního

výzkumného týmu do doby analýzy v programu R. V tomto bodě muselo dojít k několika malým úpravám, které jsou zpracovány v kapitole Analýzy v programu R.

Výsledky bakalářské práce odpovídaly výsledkům replikovaného článku. Číselné hodnoty byly totožné a až na jednu výjimku, kdy analýza bakalářské práce ukázala, že profese guard se vyskytuje z hlediska efektů mezi 6 nejsilnějšími profesemi ve 4 ze 7 korpusů a v původní analýze se objevila ve 3 ze 7 korpusů (Nejsilněji asociovaná povolání s genderem napříč korpusy), nedošlo k žádným nesrovnalostem a odlišnostem.

5 Závěr

Po dokončení výzkumu je možno zkonstatovat, že genderové stereotypy o povolání v přirozeném jazyce existují a vykazují konstantní rozdíly napříč oblastmi zkoumání. Tyto implicitní asociace jsou již zakotvené v komunikaci a neliší se ve věkových kategoriích. Zároveň mezi korpusy dospělých a korpusy dětí byla prokázána vysoká kovariance. Postupem času se lehce mění zastoupení a síla stereotypů, u nejstarších korpusů převládaly asociace práce s mužským pohlavím a u středních a novějších korpusů se hodnoty více přibližovaly k genderové vyrovnanosti. To ovšem nic nemění na faktu, že muži jsou s prací spojováni více, což je i podloženo reálnými statistikami společnosti BLS, která zveřejnila výzkum o poměru mužů a žen v zaměstnanosti u jednotlivých profesí a ukázala, že v 60 % zkoumaných pracovních pozic dominují muži.

V práci jsou také patrné rozdíly mezi jednotlivými korpusy, v některých se vyskytoval vyšší počet stereotypů o povolání spojených s ženami, konkrétně v oblasti mluvy dospělých, dětské mluvy, mluvy mířené dětem a TV dospělých SimplyScripts, jinde převládali mužské stereotypy, a to v oblastech knih pro dospělé – Gutenbergu, dětských knih a TV dětí – kombinace. Nicméně mužské stereotypy byly silnější a častěji asociované s danými povoláními, proto i průměrný efekt vyšel 0,03.

Bakalářská práce zároveň ukazuje přehled nejsilnější asociovaných povolání s jednotlivým genderem. Jako nejmužnější profese se na základě WEATU ukázaly být: machinist (opravář / údržbář strojů), excavator (bagrista), agriculturalist (agrotechnik či zemědělský odborník), geologist (geolog), pilot (pilot) a inspector (inspektor). Povolání nejvíce asociovaná s ženami byla: nurse (zdravotní sestra), maid (služebná či pokojská), dressmaker (švadlena), teacher (učitelka), veterinarian (veterinářka) a waitress (servírka). Všechny tyto profese zároveň pozitivně korelovaly s reálnými statistikami.

Všechny čtyři hypotézy byly pozitivně potvrzeny a v žádné nedošlo k výrazné rozepři, dospělo se prakticky ke stejným výsledkům jako v původní studii. Jediným rozdílným výsledkem byl počet zastoupení povolání guard (strážný) mezi šesti nejsilnější asociovanými profesemi s mužským pohlavím. Autoři tvrdili, že toto povolání bylo zastoupeno třikrát mezi šesti nejsilnější asociovanými profesemi s muži, ve výzkumu bakalářské práce se objevilo čtyřikrát.

Tématem k zamyšlení je, zda je možné, že tyto implicitně zakotvené stereotypy v přirozeném jazyce ovlivňují reálnou situaci v zaměstnanosti jednotlivých povolání, protože je dokázáno, že stereotypy mohou být škodlivé a automaticky ovlivňovat jedince v percepci druhých. To ovšem tato bakalářská práce nemohla zjistit, pouze ukázala, že genderové stereotypy o povolání existují a jak silných asociací dosahují.

Také je ukázáno, že na základě zveřejněných dat je původní výzkum schopen replikace. Jedinými nedostupnými prvotními neupravenými daty byly ty z oblasti dětských

audiovizuálních médií společnosti Disney. Naštěstí se na serveru OSF vyskytují soubory, které umožnily další postup. Další problém nastal v momentu analýzy v statistickém softwaru R a chyby jsou popsány v kapitole Nesrovnalosti a nutné úpravy. Všechny nesrovnalosti byly sepsány a odeslány emailem autorům článku.

Daná bakalářská práce a její výstupy byly konzultovány se studenty psychologických oborů, kteří byli seznámeni s metody a postupy měření výsledků, ti byli na pochybách, zda metoda fastText a jiné metody použité pro vytvoření práce skutečně měří genderové stereotypy. Dle jejich názorů správné detekování genderových stereotypů by probíhalo asi tak, že by se vybraly náhodné věty obsahující zmínky o povolání a také slova jako muž či žena (či jejich synonyma – dívka, chlapec, matka, otec, slečna, hoch, ...) z dostupných dat přirozeného jazyka (z jednotlivých neupravených korpusů například). Následně by se zaměstnal náhodný vzorek lidí, kteří by nezávisle na sobě ohodnotili dané věty a jak moc je muž nebo žena opravdu spojen s daným povoláním v náhodném vzorku zvolených vět. Stanovila by se hodnotící škála, pomocí které by lidé věty a spojení hodnotili a z těchto hodnocení by se provedla analýza T test.

6 Reference

1. Anderson, S. F., & Maxwell, S. E. (2017, March). Addressing the “Replication Crisis”: Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivariate Behavioral Research*, 52(3), 305–324. doi:10.1080/00273171.2017.1289361
2. Beyer, S. (2016, October). Low Awareness of Occupational Segregation and the Gender Pay Gap: No Changes Over a 16-year Span. *Current Psychology*, 37(1), 373–389. doi:10.1007/s12144-016-9521-4
3. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017, April). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. doi:10.1126/science.aal4230
4. Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011, June). Math-gender stereotypes in elementary school children. *Child Development*, 82, 766–779. doi:10.1111/j.1467- 8624.2010.01529.x
5. Drummond, C. (2017, December). Reproducible research: a minority opinion. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(1), 1–11. doi:10.1080/0952813x.2017.1413140
6. Dupas P., Modestino A. S., Niederle M., & Wolfers J. (2021, May). Gender and The Dynamics of Economic Seminars. *Stanford University*. Dostupné z: <https://web.stanford.edu/~pdupas/Gender&SeminarDynamics.pdf>
7. Fiani, C. N., & Han, H. J. (2018, March). Navigating identity: Experiences of binary and non-binary transgender and gender non-conforming (TGNC) adults. *International Journal of Transgenderism*. doi: 10.1080/15532739.2018.1426074
8. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644. doi:10.1073/pnas.1720347115
9. Gentzkow, M., Kelly, B., & Taddy, M. (2019, September). Text as Data. *Journal of Economic Literature*, 57(3), 535–574. doi: 10.1257/jel.20181020
10. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques, third edition (3rd ed.)*. San Francisco, Spojené státy americké: Morgan Kaufmann Publishers.
11. Hao, K. (2018). What Is Machine Learning?. *MIT Technology Review*. Dostupné z: <https://www.technologyreview.com/2016/08/11/158359/the-big-deal-with-big-data-isnt-just-the-data/>
12. Hendriks, F., Kienhues, D., & Bromme, R. (2020). Replication crisis = trust crisis? The effect of successful vs failed replications on laypeople’s trust in researchers and

- research. *Public Understanding of Science*, 096366252090238.
doi:10.1177/0963662520902383
13. Hnilica, K. (2010). *Stereotypy, předsudky, diskriminace: (pojmy, měření, teorie)*. Praha, Česko: Karolinum. Dostupné z: <https://ndk.cz/uuid/uuid:fac6af10-3997-11e8-b52f-5ef3fc9ae867>
 14. Holmes, J., Meyerhoff, M., & Blackwell Reference Online (Online service). (2005). *The handbook of language and gender*. Oxford, Anglie: Malden, MA: Blackwell Pub.
 15. Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2020, December 26). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of 65+ million words. *OSF HOME*. Dostupné z: <https://osf.io/kqux5/>
 16. Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021, February). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, 32(2), 218–240. doi: 10.1177/0956797620963619
 17. Clarivate. (2021). Psychological Science. *InCites Journal Citation Reports*. Dostupné z: <https://journalprofile-clarivate-com.zdroje.vse.cz/jif/home/?journal=PSYCHOL%20SCI&year=2019&editions=SSCI&pssid=H1-Zjlx2FgRRCx2BNLV9V7pQCTFICX6fUe9STEO-18x2dHiMmx2Bc2lGkx2Fsoix2BefrHXeAx3Dx3Dx2FnMQmWf3v8yxxODI833bD4gx3Dx3D-qBgNuLRjcgZrPm66fhjx2Fmwx3Dx3D-h9tQNJ9Nv4eh45yLvkdX3gx3Dx3D>
 18. Kitchin, R., & McArdle, G. (2016, January). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. doi:10.1177/2053951716631130
 19. Koldinská, K. (2010). *Gender a sociální právo: rovnost mezi muži a ženami v sociálněprávních souvislostech*. Praha, Česko: C.H. Beck, Dostupné z: <https://ndk.cz/uuid/uuid:9c425993-bd1f-417c-bf29-154153efb533>
 20. Kowalski, R. (1979, July). Algorithm = Logic + Control. *Communications of the ACM*. 22(7). 424–436. doi: 10.1145/359131.359136
 21. Lewis, M., & Lupyan, G. (2019, March). Gender stereotypes are reflected in the distributional structure of 25 languages. *PsyArXiv*. doi: 10.31234/osf.io/7qd3g
 22. Michalopoulos, S., & Xue, M. M. (2021, January). Folklore. *The Quarterly Journal of Economics*. doi: 10.1093/qje/qjab003
 23. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Cornell University arXiv*. Dostupné z: <https://arxiv.org/pdf/1301.3781.pdf>

24. MIT Experimental Learning In Collaboration with GetSmarter. (2016, August 11). The Big Deal with Big Data Isn't (Just) the Data. *MIT Technology Review*. Dostupné z: <https://www.technologyreview.com/2016/08/11/158359/the-big-deal-with-big-data-isnt-just-the-data/>
25. Mullainathan, S., & Spiess, J. (2017, February). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87–106. doi:10.1257/jep.31.2.87
26. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *OSF*. Dostupné z: <https://osf.io/82fth/>
27. Project Implicit. (2011). About Us. *Project Implicit*. Dostupné z: <https://implicit.harvard.edu/implicit/aboutus.html>
28. Rong, X. (2016). word2vec Parameter Learning Explained. Dostupné z: <https://arxiv.org/abs/1411.2738>
29. Schulz, D., & Bahník, Š. (2019, January). Gender associations in the twentieth-century English-language literature. *PsyArXiv Preprints*. doi: 10.31234/osf.io/grn8t
30. Tatsioni, A., & Ioannidis, J. P. A. (2017, April). Meta-analysis. *International Encyclopedia of Public Health*, 117–124. doi: 10.1016/b978-0-12-803678-5.00291-5
31. U.S. Bureau of Labor Statistics. (1998). Labor force statistics from the current population survey: 1995–1999 annual averages - household data - tables from employment and earnings (Table 10). *U.S. Bureau of Labor Statistics*. Dostupné z: https://www.bls.gov/cps/cps_aa1995_1999.htm
32. U.S. Bureau of Labor Statistics. (2019). Table A-1. Time spent in detailed primary activities and percent of the civilian population engaging in each activity, averages per day by sex, 2019 annual averages — Continued. *U.S. Bureau of Labor Statistics*. Dostupné z: www.bls.gov/tus/a1-2019.pdf
33. Vávra, M. (2015). Replikační krize v sociálních vědách a možný příspěvek datových archivů k jejímu řešení. *Data a výzkum – SDA Info*. 9(2), 121-124. Dostupné z: http://dav.soc.cas.cz/uploads/2899de6d6f4524c146ace9269f506fc23d5ea23e_DaV15_02_3recenze23.indd.pdf
34. Wessels, B., Finn, R., Wadhwa, K., Sveinsdottir, T., Bigagli, L., Nativi, S., & Noorman, M. (2017). *Open Data and the Knowledge Society*. Nizozemsko, Amsterdam: University Press. Dostupné z: <https://library.oapen.org/bitstream/id/182859cf-f9d7-4136-8413-a862804a72ed/625332.pdf>