Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir og Anton Karl Ingason

Vondar vélþýðingar

Um kynjahalla í íslenskum þýðingum Google Translate

1. Inngangur

Talað er um tæknihalla (e. *machine bias*) þegar tölvukerfi sem þjálfuð eru á raunverulegum málgögnum fara óumbeðin að endurspegla samfélagslegan ójöfnuð á borð við kynja- og kynþáttahalla óháð ásetningi þeirra sem búa kerfin til.[1] Þetta getur til að mynda gerst í vélþýðingarkerfum og öðrum máltæknilausnum sem byggja á málheildum. Það stafar af því að þau tileinka sér ekki aðeins mynstur og formgerð tungumálsins sem gögnin eru skrifuð á heldur einnig samfélags- og menningarleg fyrirbæri sem koma þar fram. Þetta getur haft í för með sér að tæknin viðhaldi og ýti undir ójöfnuð þvert á samtímalegar framfarir. Með sífelldri framþróun tækninnar og mikilvægi hennar innan samfélaga er nauðsynlegt að vera á varðbergi gagnvart slíkum réttlætisbresti. Tæknihalli er nú orðinn að viðfangsefni rannsókna innan ýmissa fræðasviða og er máltækni þar ekki undanskilin.[2]

[1] Höfundar vilja þakka ritstjórum og tveimur ónafngreindum ritrýnum fyrir gagnlegar ábendingar við gerð þessarar greinar.

[2] Sjá til dæmis Emily Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell, „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?“, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* 2021, bls. 610–623; Marcelo O. R. Prates, Pedro H. Avelar og Luís C. Lamb, „Assessing gender bias in machine translation: A case study with Google Translate“, *Neural Computing and Applications* 32/2019, bls. 6363–6381; Rachel Rudinger, Jason Naradowsky, Brian Leonard og Benjamin Van Durme, „Gender Bias in Coreference Resolution“, *Proceedings of the 2018 Conference of the*

Ritið
3. tbl. 21. árg. 2021 (177-200)
Ritrýnd grein

Útgefandi:
Hugvísindastofnun Háskóla Íslands,
Sæmundargötu 2, 102 Reykjavík

Birtist á vefnum http://www.ritid.hi.is.
Tengiliður: ritid@hi.is

DOI: 10.33112/ritid.21.3.7

Í þessari grein verður fjallað um niðurstöður rannsóknar sem gerð var á íslenskum þýðingum úr vélþýðingarkerfinu Google Translate. Markmiðið var að kanna hvort kynjahalli kæmi fram í íslenskum þýðingum líkt og rannsóknir hafa sýnt að getur gerst í öðrum málum. Í íslensku eru þrjú málfræðileg kyn: karlkyn, kvenkyn og hvorugkyn. Kyn er málfræðileg beygingarformdeild fallorða, til að mynda hafa nafnorð fast kyn en lýsingarorð fá kyn sitt frá fallorðinu sem þau standa með eða vísa til. Persónufornöfn í þriðju persónu standa alltaf í ákveðnu kyni, *hún/hán/hann/*það, en persónufornöfn í fyrstu og annarri persónu eru eins, ég/þú, óháð kyni þess einstaklings sem þau vísa til. Ef lýsingarorð á við persónufornafn í fyrstu eða annarri persónu hlýtur það kynbeygingu sína út frá samhengi. Hins vegar er þessu öðruvísi háttað í ensku þar sem hvorki nafnorð né lýsingarorð eru flokkuð eftir kynjum. Vegna þessa breytileika milli tungumálanna er unnt að kanna hvort kynjahalli komi fram þegar vélþýðingarkerfi er látið þýða setningar í fyrstu persónu með lýsingarorði frá ensku yfir á íslensku.

Í rannsókninni voru lýsingarorð þýdd úr kynhlutlausum setningum á ensku, þ.e. setningum þar sem kyn kemur ekki fram, yfir á íslensku. Í ljós kom að vélþýðingarnar virðast vera nokkuð hlutdrægar þar sem það virðist háð merkingu lýsingarorðanna hvort þau fá karlkyns- eða kvenkynsbeygingu í íslenskri þýðingu. Þegar lýsingarorð sem lýsa persónuleika og persónueinkennum fólks eru skoðuð kemur í ljós að þau sem hafa jákvæða merkingu eru líklegri til þess að fá karlkynsbeygingu en kvenkynsbeygingu í íslensku þýðingunni. Þetta eru orð eins og *sterkur* (e. *strong*) og *snjall* (e. *clever*). Af lýsingarorðum með neikvæða merkingu er hins vegar líklegra að kvenkynsmynd sé notuð í þýðingunni, t.d. *heimsk* (e. *stupid*) og *veik* (e. *weak*). Þessu er aftur á móti öfugt háttað með lýsingarorð sem lýsa útliti, en orð sem venjulega eru notuð til þess að lýsa útliti á jákvæðan hátt eru frekar þýdd með kvenkynsmynd en þau neikvæðu með karlkynsmynd. Einnig er lýsingarorðið *góð/ur* (e. *good*) skoðað í samhengi við ákveðin störf, heimilisstörf annars vegar og tækni- og iðnaðarstörf hins vegar. Þar virðist það háð eðli starfanna hvort lýsingarorðið fær kvenkyn eða karlkyn. Við höfum ekki aðgang að frumgögnum Google en trúlega endurspeglar þetta með ein-

North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2/2018, bls. 8–14; Kaiji Lu, Piotr Mardziel, Fangjing Qu, Preetam Amancharla og Anupam Datta, „Gender Bias in Neutral Natural Language Processing“, *Logic, Language, and Security*, ritstj. Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo og Mitsuhiro Okada, Cham: Springer, 2020, bls. 189–202.

hverjum hætti dreifingu gagna í þeim málheildum sem liggja að baki þýðingarlíkönum kerfisins.

Niðurstöðurnar sýna að töluverður munur er á birtingarmynd karlkyns annars vegar og kvenkyns hins vegar í þýðingarforritinu Google Translate og má þar sjá mynstur sem samsvara ákveðnum samfélagslegum hugmyndum um kyn og kynjahlutverk. Almennt er meiri og víðtækari umræða um karlmenn í samfélaginu en konur. Sú orðræða sem beinist að konum einblínir hins vegar oft á útlit þeirra fremur en til dæmis afrek í starfi eða jákvæða eiginleika í fari þeirra. Einnig hafa kynjahlutverk verið samofin menningunni í gegnum tíðina þar sem heimilisstörf hafa verið ætluð konum og iðnaðarstörf körlum. Það er skýrt að mynstrið sem kemur fram í þýðingunum endurspeglar þennan samfélagslega kynjahalla.

Efnisskipan greinarinnar er á þessa leið: Í 2. kafla verður fjallað um vélþýðingar og málheildir, greint frá þeim aðferðum sem notaðar eru í slíkum kerfum og sjónum beint að vélþýðingarforritinu Google Translate. Sagt verður frá rannsóknum á tæknihalla og sérstaklega einblínt á kynjahalla. Í 3. kafla verður fjallað um málfræðilegt kyn og mismunandi kynjaaðgreiningu í íslensku og ensku. Í 4. kafla verður aðferðafræði rannsóknarinnar lýst og í 5. kafla verður greint frá tölulegum niðurstöðum hennar og dæmi sýnd um þýðingar. Í 6. kafla eru umræður þar sem niðurstöður rannsóknarinnar eru túlkaðar og settar í samhengi við samfélagslegar hugmyndir um kyn og kynjahlutverk. Þá eru lokaorð í 7. kafla.

2. Vélþýðingar

Þegar tölvur eru notaðar til að þýða texta sjálfvirkt af einu tungumáli yfir á annað er talað um vélþýðingar (e. *machine translation*). Fyrstu vélþýðingarkerfin, sem komu fram um miðja 20. öld, byggðust á reglum sem smíðaðar voru fyrir hvert tungumálapar fyrir sig. Á níunda áratugnum varð svo mikil gróska í vélþýðingum og ný kerfi sem byggðu m.a. á dæmum og tölfræði urðu meira áberandi.[3] Slík kerfi eru enn í dag mikið notuð en í stað reglna byggja þau á miklu magni af raunverulegum málgögnum, þ.e. samansafni af textum sem skrifaðir eru af fólki. Hefðbundin tölfræðileg vélþýðingarkerfi nota svokölluð forsagnarlíkön (e. *prescriptive model*) til að kenna tölvum að þýða á milli tungumála. Það gera þau með því að greina stórar samhliða

[3] Stephen DellaPietra og Vincent DellaPietra, „Candide: A statistical machine translation system“, *Proceedings of the workshop on Human Language Technology*, 1994, bls. 457.

málheildir (e. *parallel corpora*), þ.e. texta sem skrifaðir eru á einu tungumáli og hafa verið þýddir af mannlegum þýðendum yfir á annað. Textanum er þá ýmist skipt niður í stök orð, setningar eða setningahluta og kerfið vinnur úr málheildinni upplýsingar um hvernig þessar máleiningar eru þýddar og líkindi á tilteknum þýðingum. Þessar upplýsingar eru svo notaðar til að leiða út þýðingu á texta sem sendur er inn í kerfið.[4]

Samhliða málheildir sem þýðingarkerfi byggja á samanstanda af alls kyns textasöfnum sem til eru á báðum tungumálum tungumálaparanna, þ.e. textum sem hafa verið þýddir af mannlegum þýðendum. Þessi gagnasöfn þurfa að vera býsna stór til þess að vélþýðingakerfi geti náð góðum árangri. Æskilegt er að þau innihaldi 25–35 milljón setningapör, en talið er að smærri málheildir geti þó náð ágætum árangri innihaldi þær að minnsta kosti 2 milljónir setningapara. Sem dæmi má nefna ensk-íslensku samhliða málheildina *ParIce* sem í fyrstu útgáfu inniheldur 3,5 milljónir setningapara. Hún er gerð með það að markmiði að verða nægilega stór og vönduð svo hægt sé að nýta hana til þjálfunar á vélþýðingarkerfi með árangursríkum hætti. Hún samanstendur af 11 textasöfnum, þá aðallega af opinberum skjölum Evrópska efnahagssvæðisins og skjátextaþýðingum.[5]

Eitt þekktasta og mest notaða vélþýðingarkerfið í dag er Google Translate, þýðingarveita tæknirisans Google. Hún er ókeypis og aðgengileg hverjum sem er í gegnum netið og notendur eru yfir 500 milljónir.[6] Google Translate var fyrst gefið út árið 2006 og þýddi þá aðeins á milli tveggja tungumála, arabísku og ensku.[7] Síðan þá hefur þýðingarvélin þróast og stækkað og styður nú 108 tungumál.[8] Upphaflega var enska eins konar milliliður, þ.e. þegar sleginn var inn texti á frummáli var fyrsta skrefið að þýða hann yfir á ensku áður en hann var þýddur yfir á markmálið. Þetta þótti fýsilegasti

[4] Anna Björk Nikulásdóttir, Jón Guðnason og Steinþór Steingrímsson, *Máltækni fyrir íslensku 2018–2022: Verkáætlun*, Reykjavík: Mennta- og menningarmálaráðuneytið, 2017, bls. 72.

[5] Starkaður Barkarson og Steinþór Steingrímsson, „Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus“, *Proceedings of the 22nd Nordic Conference on Computational Linguistics* 2019, bls. 140–145; Starkaður Barkarson og Steinþór Steingrímsson, „ParIce: English-Icelandic parallel corpus“, *CLARIN-IS*, http://hdl.handle.net/20.500.12537/16.

[6] Barak Turovsky, „Ten years of Google Translate“, *Google*, 2016, sótt 12. apríl 2021 af https://www.blog.google/products/translate/ten-years-of-google-translate/.

[7] Franz Och, „Statistical machine translation live“, *Google AI Blog*, 2006, sótt 12. apríl 2021 af https://ai.googleblog.com/2006/04/statistical-machine-translation-live.html.

[8] Google Translate, sótt 12. apríl 2021 af https://translate.google.com/intl/en/about/languages/.

kosturinn þar sem óraunhæft var að ráða mennska þýðendur fyrir hvert mál og hefði í raun verið tilgangslaust, meðal annars vegna þess hversu hratt tungumál breytast.[9] Samhliða málheildin, sem Google Translate byggði þá á, samanstóð af opinberum textum úr skjalasafni Sameinuðu þjóðanna og Evrópuþingsins, sem voru þegar þýddir af mannlegum þýðendum yfir á um 30 tungumál.[10] Google Translate byggði upphaflega aðeins á tölfræðilegu vélþýðingarkerfi sem reiknar út hvaða þýðing er tölfræðilega líklegust. Árið 2016 tók þýðingarvélin svo að nota kerfi sem byggir á tauganeti.[11]

Tauganet (e. *neural machine translation*) eru nýjasta aðferðin í vélþýðingum og slíkar þýðingarvélar ná betri árangri en hefðbundnar tölfræðilegar þýðingavélar. Líkt og tölfræðilegar þýðingavélar byggja tauganet á raunverulegum málgögnum úr samhliða málheildum en þau nýta sér einnig einmála málheildir til þess að læra betur hvernig setningar á markmálinu eru myndaðar. Í stað tölfræðilegra reikninga lærir tauganetið mynstur og duldar formgerðir í málinu sem það finnur í gögnunum. Þessi mynstur geta til dæmis verið málfræði tungumálanna og undantekningar frá málfræðireglum eða varpanir á milli orðmynda, orða eða orðasambanda úr einu tungumáli yfir á annað. Þannig „lærir“ þýðingarvélin tungumálin sem um ræðir og textinn sem hún skilar verður eðlilegri, þ.e. hún er líklegri til að skila texta sem líkist mannlegu máli.[12] Óhjákvæmilega lærir tauganetið þó ekki aðeins tungumálið sjálft, þ.e. orðin og málfræðina, heldur einnig samfélagsleg mynstur sem koma fram í orðræðunni.

2.1 Tæknihalli

Málheildir sem vélþýðingar og önnur máltækni byggja á innihalda texta sem allir eru skrifaðir af raunverulegu fólki. Því er eðlilegt að í þeim komi

[9] Sjá t.d. Marcelo O. R. Prates o.fl., „Assessing gender bias in machine translation: A case study with Google Translate“, bls. 6378; Christian Boitet, Mark Seligman, Hervé Blanchon og Valérie Bellynck, „MT on and for the Web“, *Proceedings of the 6th International Conference on Natural Language Proccessing and Knowledge Engineering*, Peking, Kína, 2010, bls. 1–10, hér bls. 6; Martin Benjamin, „How GT Pivots through English“, *Teach You Backwards*, 2019, sótt 12. apríl 2021 af https://www.teachyoubackwards.com/extras/pivot/.

[10] Marcelo O. R. Prates o.fl., „Assessing gender bias in machine translation: A case study with Google Translate“, bls. 6366.

[11] Barak Turovsky, „Found in translation: More accurate, fluent sentences in Google Translate“, *Google*, 2016, sótt 12. apríl 2021 af https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.

[12] Anna Björk Nikulásdóttir o.fl., *Máltækni fyrir íslensku 2018–2022: Verkáætlun*, bls. 72-73.

fram alls kyns samfélagslegar hugmyndir og mynstur sem menningin hefur mótað og jafnframt litað tungumálin í gegnum tíðina. Þar sem tölfræðileg vélþýðingarkerfi og þýðingarvélar sem byggja á tauganeti læra að þýða milli tungumála með því að greina þessi raunverulegu málgögn getur verið hætta á því að þau endurspegli samfélagslega vankanta líkt og fordóma gagnvart kyni og kynþáttum.[13]

Vélrænt nám (e. *machine learning*) á borð við það sem fer fram í tölfræðilegu þýðingarkerfi á sér einnig stað á öðrum sviðum tækninnar og þar hafa komið fram dæmi um tæknihalla sem sýnir ómeðvitaða hlutdrægni. Sem dæmi gerðist það hjá tölfræðilegu myndgreiningarkerfi Google Photos að myndir af þeldökku fólki voru merktar sem myndir af górillum.[14] Einnig hafa komið fram dæmi um það að andlitsgreining sem notuð er til þess að aflæsa símunum iPhone X frá Apple hafi ekki getað greint á milli tveggja einstaklinga af asískum uppruna. Enn annað dæmi um slíkt vélrænt nám er að tækni sem notuð hefur verið til þess að reikna og spá fyrir um glæpahegðun fólks í Bandaríkjunum hefur sýnt neikvæða hlutdrægni gagnvart þeldökkum föngum.[15]

Nýleg grein eftir Emily Bender o.fl.[16] kafar djúpt í það hvernig kynjahalli og annars konar félagsleg bjögun birtist í vélgerðum textum. Hún skoðar ítarlega nokkur mállíkön (e. *language model*) sem þjálfuð eru á ýmsum málheildum og veltir því upp hvort slíkar málheildir geti í raun verið „of stórar". Því stærri sem málheildir eru því víðtækari er uppruni textanna. Sem dæmi er netið algengur vettvangur til þess að nálgast mikið magn af skrifuðu efni frá almenningi en ákjósanlegt er að málheildir innihaldi sem fjölbreyttastar textategundir úr ólíkum áttum. Samt sem áður geta falist í þessu ákveðnar leyndar hættur. Þegar efni er til að mynda safnað af samfélagsmiðlum og öðrum opnum og aðgengilegum síðum (svo sem Twitter, Reddit eða Wikipedia) þarf að hafa í huga hverjir notendur þeirra eru, með tilliti til félagslegra þátta eins og stéttaskiptingar, kynþáttar, aldurs og menningarlegs umhverfis. Bender o.fl.[17] vísa í þessu samhengi í rannsóknir sem sýna

[13] Keith Kirkpatrick, „Battling Algorithmic Bias: How do we ensure algorithms treat us fairly?", *Communications of the ACM*, 2016, bls. 16–17, hér bls. 16.

[14] Megan García, „Racist in the Machine: The disturbing Implications of Algorithmic Bias", *World Policy Journal* 33: 4/2016–2017, bls. 111–117.

[15] Marcelo O. R. Prates o.fl., „Assessing gender bias in machine translation: A case study with Google Translate", bls. 6364.

[16] Emily Bender o.fl., „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"

[17] Emily Bender o.fl., „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", bls. 613.

fram á að notendur og höfundar texta á þessum miðlum séu að stórum hluta hvítir ungir karlmenn úr vestrænu samfélagi. Þetta leiðir til þess að textar frá jaðarsettum hópum eru í minnihluta og hætta á að sjónarhornið sem kemur fram í málheildunum takmarkist við ríkjandi hugmyndir ákveðinna forréttindahópa. Því er hætta á að fordómar, hlutdrægni og hatursorðræða frá takmörkuðu sjónarhorni verði hluti af gögnunum og endurspeglist í mállíkönum sem þjálfuð eru á þeim.[18] Þetta getur valdið því að jákvæð samfélagsleg þróun, svo sem hugmyndir um jafnrétti kynjanna, skili sér ekki í úttaki máltækniafurða.

Þýðingavélar geta reynst ágætt verkfæri til þess að athuga hvort og hvernig kynjahalli (*e. gender bias*) kemur fram í máltækni. Rannsóknir á kynjahalla í vélþýðingum hafa komið fram á sjónarsviðið í auknum mæli undanfarið og sýnt að slíkur halli sé til staðar. Sem dæmi má nefna rannsóknir Londu Schiebinger á birtingarmynd kynjanna í vísindum, sem vörpuðu ljósi á hversu mikið hallar á konur í vísindalegu samhengi.[19] Ein þessara rannsókna tók til Google Translate en þar kom í ljós að þýðingarvélin birti sjálfkrafa karlkyns persónufornöfn þar sem þau áttu ekki við. Schiebinger nefnir sem dæmi grein um vísindakonu sem skrifuð var á spænsku, þar sem frumlagsfornöfn eru ekki nauðsynleg í setningagerðinni. Þegar greinin var þýdd með Google Translate yfir á ensku, þar sem fornöfn eru nauðsyn, reiknaði forritið út að karlkynsfornafnið væri líklegast. Þetta segir Schiebinger stafa af því að mun oftar er vísað til karlkyns í skrifuðum heimildum en kvenkyns og vísar þar í Google NGram málheildina.[20] Þetta sýnir fram á að kynjahalli getur verið vandamál í vélþýðingum og Schiebinger bendir á að þannig hafi félagslegar hugmyndir fortíðarinnar ósjálfrátt áhrif á ójöfnuð í framtíðinni. Þess vegna sé mikilvægt að vera meðvituð um og bregðast við vandamálinu strax við gerð mállíkansins.[21]

Ljóst er að gerð mállíkana sem byggja á málheildum er margþætt og flókið verkefni. Við þróun tækninnar er eðlilegt að fyrst sé áhersla lögð á

[18] Sama heimild, bls. 613–615.

[19] Cara Tannenbaum, Robert P. Ellis, Friederike Eyssel, James Zou og Londa Schiebinger, „Sex and gender analysis improves science and engineering“, *Nature* 575/2019, bls. 137–146.

[20] Londa Schiebinger, „Gendered Innovations: Harnessing the Creative Power of Gender Analysis“, *Feature new perspectives* 2016, bls. 28–31, hér bls. 29; Cara Tannenbaum o.fl., „Sex and gender analysis improves science and engineering“, bls. 140.

[21] Londa Schiebinger, „Gendered Innovations: Harnessing the Creative Power of Gender Analysis“, bls. 30.

árangur tólanna og nytsemi þeirra, sem í tilfelli vélþýðinga felst í því að þýða texta rétt á milli tungumála svo niðurstaðan verði sem líkust eðlilegu máli. Spurningin er hins vegar hvort þeir sem þróa búnaðinn séu ómeðvitaðir um að tæknihalli komi fram í honum eða hvort þeir sjái það sem ásættanlegan fórnarkostnað til að ná meiri gæðum á öðrum sviðum í þýðingunni. Það ætti að vera markmið þeirra að tæknin fylgi jákvæðri framþróun á öllum sviðum, en valdi því ekki að samfélagsbreytingar stöðvist eða gangi til baka. Þess vegna er mikilvægt að rannsaka og benda á það þegar tæknin hallar á ákveðna hópa samfélagsins.

Annað dæmi um rannsókn á kynjahalla í vélþýðingum er tilraun Prates o.fl.[22] sem leiddi í ljós að ákveðinn kynjahalli væri til staðar í Google Translate með því að rannsaka hvaða kyn voru tengd við mismunandi starfsheiti. Þetta var gert með því að þýða setningar úr tólf tungumálum sem hafa ekki málfræðilega kynjaaðgreiningu, til dæmis ungversku, yfir á ensku þar sem kynjaaðgreining kemur aðeins fram í fornöfnum í þriðju persónu, en þau hafa þrjú mismunandi kyn. Setningarnar á frummálinu innihéldu kynhlutlaus persónufornöfn og vísuðu til ákveðinna starfsheita. Kynjahallinn kom þá fram í því að það virtist háð eðli starfanna hvort enska þýðingin fengi karlkyns eða kvenkyns persónufornafn eins og sjá má í (1).

(1)	Ungverska		Enska
	ő egy ápolónő	**>**	**she is a nurse**
	hann/hún/það er hjúkrunarfræðingur		hún (kvk) er hjúkrunarfræðingur
	ő egy mérnök	**>**	**he is an engineer**
	hann/hún/það er verkfræðingur		hann (kk) er verkfræðingur

Niðurstöður Prates o.fl.[23] sýndu að karlkyns persónufornöfn voru almennt mun líklegri til þess að koma fram í ensku þýðingunum en kvenkyns eða hvorugkyns fornöfn. Þar kom fram skýr munur milli kynjanna þegar störfin voru flokkuð saman eftir því hvert eðli þeirra var og á hvaða sviði atvinnulífsins þau voru. Karlkyn var meira áberandi með störfum þar sem staðalímyndin er sú að það séu aðallega karlar sem sinni þeim. Til að mynda var það aðeins í 4% tilvika sem setningar í flokki svokallaðra STEM starfa (vísindi, tækni, verkfræði og stærðfræði) fengu persónufornafn í kvenkyni. Þær fengu hins vegar persónufornafn í karlkyni í 72% tilvika. Til samanburðar var dreifingin jafnari milli kynja í flokkum heilbrigðis- og kennslustarfa þar sem

[22] Marcelo O. R. Prates o.fl., „Assessing gender bias in machine translation: A case study with Google Translate“.

[23] Sama heimild.

23% fornafna fengu kvenkynsbeygingu og 50% karlkynsbeygingu. Með því að styðjast við tölulegar upplýsingar frá stofnuninni U.S. Bureau of Labor Statistics gátu Prates o.fl. sýnt fram á að þessar niðurstöður endurspegluðu ekki hlutfallslegan mun á kyni þeirra sem unnu þessi störf í Bandaríkjunum á þeim tíma sem rannsóknin var gerð. Það er því ljóst að þetta gefur ekki rétta mynd af stöðunni eins og hún er í dag og er annað dæmi um augljósan kynjahalla í máltækni. Hann orsakast mögulega af gömlum hugmyndum um kynjahlutverk og vinnumarkaðinn sem birtast í þeim gögnum sem þýðingarvélin er þjálfuð á. Tæknin er stór hluti af nútímasamfélagi og því er mikilvægt að hún stuðli ekki að því að þessi ójöfnuður viðhaldist.

3. Kyn

Íslenska hefur þrjú málfræðileg kyn: karlkyn, kvenkyn og hvorugkyn. Nafnorð hafa fast kyn en kyn lýsingarorða aðlagast því nafnorði sem þau standa með eða vísa til. Þannig er kyn beygingarformdeild lýsingarorða.[24] Kyn nafnorða telst til orðasafnsþátta þar sem það fylgir orðinu hvar og hvenær sem er, en kyn lýsingarorða er hins vegar aðlögunarþáttur þar sem þau aðlagast því nafnorði sem þau standa með hverju sinni.[25] Kyn lýsingarorða ákvarðast því af stöðu þeirra innan setninga.

Kyn persónufornafna ræðst líka af samhengi. Stundum ræðst það af málfræðilegu samræmi við annað orð og stundum af merkingarlegu kyni. Í aðgreiningu persónufornafna er gerður greinarmunur á því hvort vísað er til talanda, viðmælanda eða einhvers annars. Sú beygingarformdeild er kölluð persóna, nánar tiltekið fyrsta, önnur og þriðja persóna. Þegar vísað er til einstaklinga með persónufornafni í þriðju persónu fer það eftir merkingarlegu kyni þeirra hvaða persónufornafn er notað, *hún/hán/hann/það*. Í fyrstu og annarri persónu, ég/þú, er hins vegar ekki gerður greinarmunur á kyni, þ.e. orðin eru eins óháð því af hvaða kyni talandi eða viðmælandi er. Það er því ekki augljóst hvert merkingarlegt kyn einstaklingsins er þegar vísað er til hans með persónufornöfnum í fyrstu og annarri persónu. Það þarf að skýrast af samhengi.

Nútímaenska hefur ekki mikla málfræðilega kynjaaðgreiningu samanborið við íslensku. Fornenska hafði þrjú málfræðileg kyn en kyn er ekki lengur beygingarþáttur í ensku þótt leifar af því megi finna í persónufornöfnum í þriðju persónu, eintölu. Líkt og í íslensku fer það eftir kyni ein-

[24] Guðrún Kvaran, *Íslensk tunga 2: Orð*, Reykjavík: Almenna bókafélagið, 2005, bls. 173–174.

[25] Sama heimild, bls. 190.

staklings hvaða fornafn er notað þegar vísað er til hans í þriðju persónu, *he/she*. Kyn fornafnanna hefur þó ekki áhrif á önnur orð í setningunni, eins og lýsingarorð sem haldast óbreytt. Í setningum með fornafni í fyrstu og annarri persónu er ekkert sem segir til um kyn einstaklingsins sem vísað er til. Því þarf samhengi til þess að merkingin komist til skila.[26]

Það liggur því beint við að kanna kynjahalla í vélþýðingum milli ensku og íslensku með því að nota setningar með fornafni í fyrstu persónu og lýsingarorði. Slíkar setningar eru kynlausar í ensku en kynjaðar í íslensku á þann veg að lýsingarorð þurfa að kynbeygjast. Við þær aðstæður þarf hugbúnaðurinn að ákveða hvaða kyn lýsingarorðið á að fá en það gerir hann með því að reikna út hvaða þýðing er tölfræðilega líklegust til að vera rétt. Þetta er sambærilegt aðferðum Prates o.fl.[27] sem nýttu sér þennan breytileika milli tungumála til að kanna kynjahalla í vélþýðingum.

Í þessu samhengi er rétt að taka fram að hvorugkynsbeyging er ekki til rannsóknar hér. Það er sökum þess að í íslensku er venjan ekki sú að hvorugkynsmyndir lýsingarorða séu notaðar til þess að lýsa fólki. Fyrir því eru sögulegar ástæður en í frumindóevrópsku voru aðeins tvö kyn, annað var notað yfir lifandi verur og hitt um dauða hluti. Hvorugkyn í íslensku er talið beintengt því síðarnefnda.[28] Það er því hægt að draga þá ályktun að íslenskum málhöfum þyki óeðlilegt að tala um manneskjur með hvorugkynsbeygingu. Í íslensku samfélagi hefur þó orðið mikil vitundarvakning varðandi fjölbreytileika kynjanna og er svokallað *kynhlutlaust mál* að ryðja sér til rúms. Það felur m.a. í sér að fólk sem skilgreinir sig ekki innan tvíkynjakerfisins, þ.e. kynsegin fólk, kýs sumt að til þess sé vísað með hvorugkyni.[29] Þessi breyting er þó of stutt á veg komin til þess að hægt sé að gera ráð fyrir að hún sé orðin hluti af náttúrulegum málgögnum sem vélþýðingarkerfi eins og Google Translate byggja á.

[26] Rodney Huddleston og Geoffrey K. Pullum, *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press, 2002.

[27] Marcelo O. R. Prates o.fl., „Assessing gender bias in machine translation: A case study with Google Translate“.

[28] Jón Axel Harðarson, „Um karla og karlynjur“, Íslenskt mál *og almenn málfræði* 23/2001, bls. 253–274, hér bls. 254, 262.

[29] Sjá t.d. Selmu Margréti Sverrisdóttur, „Hann, hún og það... eða hvað? Um kynhlutlausa persónufornafnið hán“, BA-ritgerð í almennum málvísindum við Háskóla Íslands, 2016, http://hdl.handle.net/1946/24447; Alda Villiljós, „Hán – nýtt persónufornafn?“, *Knúz – femínískt vefrit*, sótt 12. apríl 2021 af https://knuz.wordpress.com/2013/09/09/han-nytt-personufornafn/; Hildur Lilliendahl Viggósdóttir, „Annar ófrískur, hinn á túr: Tilraunir til breytinga á tungumálinu í þágu jafnréttis“, BA-ritgerð í íslensku við Háskóla Íslands, 2020, http://hdl.handle.net/19463/6955.

4. Rannsókn á kynjahalla í íslenskum þýðingum Google Translate

Rannsóknin sem hér verður fjallað um var gerð vorið 2020. Markmið hennar var að athuga hvort svipaður kynjahalli og sagt var frá í 2. kafla kæmi fram í íslenskum þýðingum Google Translate. Þetta var gert með því að slá inn í kerfið setningar á ensku sem innihalda orð sem hafa ekki málfræðilegt kyn og láta þýða þær yfir á íslensku þar sem orðin verða að standa í tilteknu kyni. Þetta voru setningar sem innihéldu persónufornafn í fyrstu persónu með lýsingarorði sem vísar til einstaklingsins sem persónufornafnið á við. Þegar vísað er til einstaklings í fyrstu persónu er ekkert í ensku formgerðinni sem segir til um hvert merkingarlegt kyn hans er. Því þarf vélþýðingarkerfið að reikna út hvaða kyn lýsingarorðið á að fá í íslensku þýðingunni. Það gerir kerfið með því að nota þýðingarlíkan sem byggir á málheildum og finna þá þýðingu sem er tölfræðilega líklegust.

Rannsakendur völdu samtals 328 lýsingarorð til þess að nota í könnuninni.[30] Þeim var skipt í tvo meginflokka, þ.e. lýsingarorð sem lýsa persónuleika og persónueinkennum og lýsingarorð sem lýsa útliti. Setningarnar sem slegnar voru inn í kerfið höfðu allar sömu formgerð á ensku sem sjá má í dæmi (2).

(2)	**Enska**	I	am	strong/weak/beautiful/ugly
		pfn.1p.	so.	lo.
	Íslenska	Ég	er	sterk-ur/veik-ur/falleg-ur/ljót-ur
		pfn.1p.	so.	lo.**kk/kvk**

Setningarnar innihéldu persónufornafn í fyrstu persónu, sögnina *to be* ‘að vera’ í framsöguhætti og eitt lýsingarorð sem lýsir annaðhvort persónuleika eða útliti. Dæmið sýnir að í ensku formgerðinni er engin kynjaaðgreining en hún þarf að vera til staðar í þeirri íslensku.

Þá voru einnig skoðaðar setningar sem lýsa færni í ákveðnum störfum. Þar var lýsingarorðið alltaf það sama, þ.e. *good* ‘góð/ur’, en stóð með mismunandi orðum sem tengjast annars vegar heimilisstörfum, t.d. *að elda* ‘cooking’, og hins vegar ýmsum iðnaðarstörfum, til dæmis *að steypa* ‘mortaring’. Alls voru skoðuð 36 orð sem tengdust mismunandi störfum, þar af 21 sem tengdust heimilisstörfum og 15 iðnaðarstörfum. Þetta var gert með hliðsjón

[30] Það skal tekið fram að val lýsingarorðanna takmarkast við hugarflug höfunda. Yfirlesari benti á að hér hefði verið upplagt að styðjast við Íslenskt orðanet (https://ordanet.arnastofnun.is) við val á lýsingarorðum. Höfundar taka undir það og munu hafa það í huga í framhaldsrannsóknum um sama efni.

af rannsókn Prates o.fl. (2019) og samfélagslegum hugmyndum um kynjahlutverk. Dæmi um formgerð slíkra setninga má sjá í (3).

(3)	I	am	(not)	good	at	sewing/woodworking	
	pfn.1p.et.	so.1p.et		lo.	fs.	so.	
	Ég	er	**(ekki)**	**góð/ur**	í	(að)	**sauma/trésmíði**
	pfn.1p.et.	so.1p.et		lo. (kk/kvk)	fs.	nhm.	(so./no.)

Ensku setningarnar sem slegnar voru inn í forritið innihéldu persónufornafn í fyrstu persónu, sögnina *to be* 'að vera' í framsöguhætti, lýsingarorðið *good* 'góður' og sagnorð í lýsingarhætti nútíðar sem táknar ákveðið starf eða iðju. Eins og dæmið sýnir verður lýsingarorðið í íslensku þýðingunni alltaf að fá málfræðilegt kyn. Sömu setningar, með sama sagnorði, voru svo prófaðar með neituninni *not* 'ekki' til þess að athuga hvort sú merkingarbreyting hefði áhrif á kyn lýsingarorðsins í íslensku þýðingunni.

Þýðingarnar voru skráðar niður í töflureikni ásamt setningunni á ensku og þar voru þær flokkaðar eftir því í hvaða kyni lýsingarorðið stóð og hvort merking lýsingarorðsins teldist jákvæð, neikvæð eða hlutlaus. Flokkunin á merkingu orðanna var byggð á mati rannsakenda og er ekki hægt að alhæfa að allir hafi sömu tilfinningu fyrir gildi og merkingu þeirra. Þó var leitast við að byggja matið á þekktum samfélagslegum hugmyndum um hvernig fólki er lýst.

Í heildina voru 446 orð prófuð í vélþýðingarkerfinu. Þar af þurfti að sigta frá 103 orð, 64 vegna þess að þýðingarnar voru rangar, lélegar eða formgerðin breyttist um of, og 22 lýsingarorð sem hafa eins beygingarmyndir í karlkyni og kvenkyni. Þá var heldur ekki hægt að nota 15 óbeygjanleg lýsingarorð, t.d. orð með beygingarendinguna *-andi*, og tvö lýsingarorð sem komu út í hvorugkyni. Aftur á móti voru þau lýsingarorð sem hafa eins beygingarmyndir í kvenkyni og hvorugkyni ekki útilokuð. Það er vegna áðurnefndrar málvenju að nota hvorugkynsbeygingar síður um fólk. Með því að útiloka hvorugkynsbeygingu lýsingarorðanna eru rannsakendur ekki að útiloka kynhlutlaust mál en til einföldunar verður hér aðeins litið á karlkyn og kvenkyn.

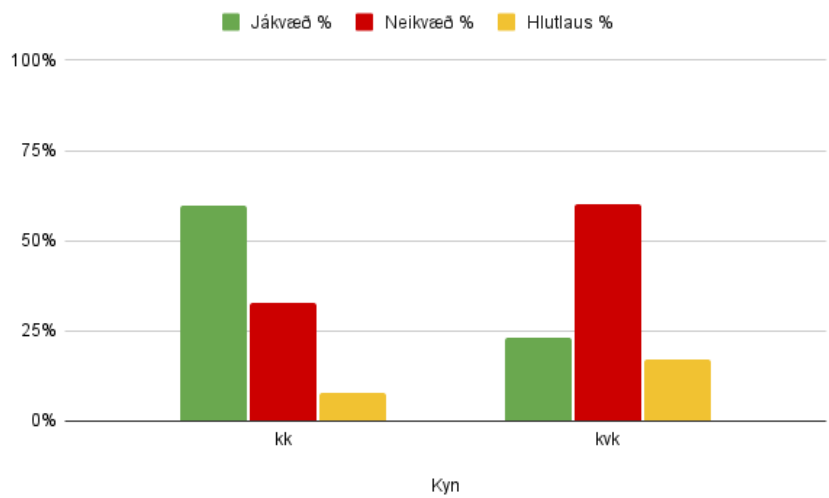
5. Niðurstöður

5.1 Persónulýsandi lýsingarorð

Lýsingarorðin sem voru rannsökuð voru sem fyrr segir flokkuð í tvo meginflokka eftir merkingu. Annar þeirra var flokkur persónulýsandi lýsingarorða,

það er lýsingarorð sem lýsa persónueinkennum fólks. Þau voru 262 talsins. Í íslensku þýðingunum frá Google Translate birtust 156 þeirra í karlkyni en 65 í kvenkyni, önnur voru ýmist í hvorugkyni, óbeygjanleg eða eins í báðum kynjum og voru því ekki talin með í samantekt gagnanna. Karlkyn er sjálfgefið kyn í íslensku fyrir mannverur í ýmsu samhengi, sem merkir að það getur vísað til allra kynja og hefur þar með víðasta notkunarsviðið.[31] Því er eðlilegt að ætla að meirihluti orðanna fái karlkynsbeygingu í íslensku þýðingunum.

Það sem er þó áhugavert við þessar niðurstöður er að það virðist vera háð merkingu orðanna hvaða kyn þau fá í íslensku þýðingunni. Jákvæð lýsingarorð, þ.e. lýsingarorð sem teljast lýsa jákvæðum eiginleikum í fari einstaklinga, voru mun líklegri til þess að fá karlkyn í íslensku en kvenkyn. Eins og sjá má á *mynd 1* voru jákvæð lýsingarorð rúm 59% af lýsingarorðum sem birtust í karlkyni á móti aðeins 23% í kvenkyni. Þessu virðist hins vegar vera öfugt háttað þegar kemur að neikvæðum lýsingarorðum. Meirihluti þeirra orða sem fengu kvenkyn í íslensku þýðingunni voru neikvæð, þ.e. þau lýsa manneskjum á neikvæðan hátt. Þetta má einnig sjá á *mynd 1* þar sem 60% lýsingarorða sem fengu kvenkynsbeygingu í þýðingunni voru neikvæð, en aðeins 32% orða sem fengu karlkyn.



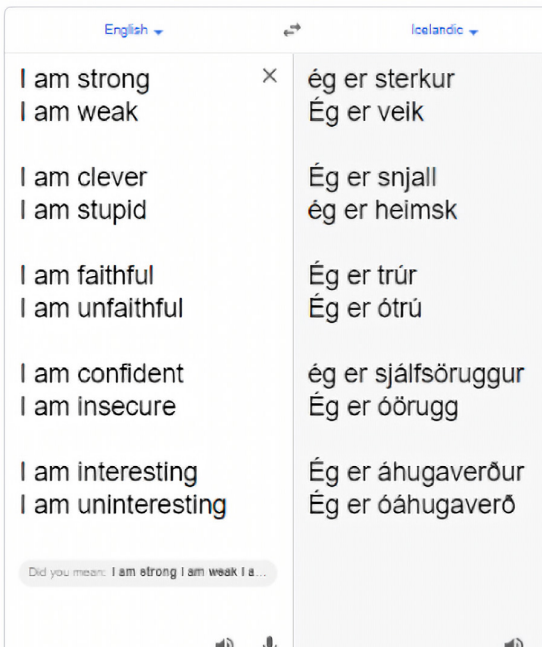
Mynd 1. Hlutfall jákvæðra og neikvæðra persónulýsandi lýsingarorða eftir kyni

[31] Höskuldur Þráinsson, *Íslensk tunga 3: Setningar*, Reykjavík: Almenna bókafélagið, 2005, bls. 83.

Þetta merkir að af þeim 65 orðum sem fengu kvenkynsbeygingu í íslenskri þýðingu voru 39 lýsingarorð sem lýsa persónueinkennum fólks á neikvæðan hátt. Þessi munur kemur skýrt fram í mörgum andstæðupörum lýsingarorða þar sem annað er talið neikvætt og hitt jákvætt. Dæmi um þetta má sjá í (4).

(4) a) Enska: I am strong_
Íslenska: Ég er **sterkur** (kk) (jákvætt)
b) Enska: I am weak
Íslenska Ég er **veik** (kvk) (neikvætt)

Hér er um andstæðupar að ræða þar sem jákvæða lýsingarorðið *strong* í (4a) fær karlkynsbeygingu í íslensku þýðingunni, *sterkur*. Aftur á móti fær andstæðan, neikvæða lýsingarorðið *weak* í (4b), kvenkynsbeygingu. *Mynd 2* sýnir skjáskot af fleiri andstæðupörum eins og þau birtast í Google Translate. Þar sést að jákvæða orðið fær karlkynsbeygingu í íslensku þýðingunni en það neikvæða fær kvenkyn.



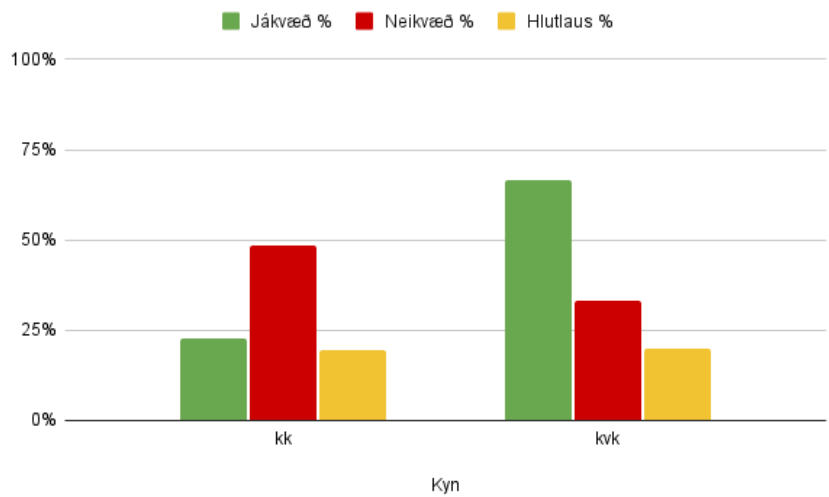
Mynd 2. Skjáskot af Google Translate

Ensku setningarnar eru allar nákvæmlega eins að gerð fyrir utan lýsingarorðin sjálf. Það er því ekkert í ensku setningunum sem gefur til kynna í

hvaða kyni lýsingarorðið á að beygjast í íslensku. Mynstrið bendir því til að það sé merking lýsingarorðsins sem stjórnar því hvaða kyn það fær, með þeim afleiðingum að orð sem vísa í neikvæð persónueinkenni fá kvenkynsbeygingu.

5.2 Útlitslýsandi lýsingarorð

Hinn flokkur lýsingarorðanna sem var kannaður er flokkur útlitslýsandi lýsingarorða, þ.e. orð sem notuð eru til þess að lýsa útliti fólks. Sá flokkur var talsvert minni en flokkur persónulýsandi orða eða 67 orð talsins. Líkt og í hinum flokknum voru mun fleiri orð sem fengu karlkynsbeygingu en kvenkyns. Lýsingarorð í karlkyni voru 31 á móti 15 í kvenkyni. Niðurstöðurnar í þessum flokki voru þvert á niðurstöður hins flokksins þegar litið er til þess hvernig kynbeygingin skiptist milli jákvæðra og neikvæðra lýsingarorða. Þetta má sjá á *mynd 3*.



Mynd 3. Hlutfall jákvæðra og neikvæðra útlitslýsingarorða eftir kyni

Af þeim lýsingarorðum sem fengu kvenkynsbeygingu hafði meirihlutinn jákvæða merkingu, þ.e. þau lýsa útliti á jákvæðan hátt. Þannig voru 10 orð af 15 með kvenkynsbeygingu jákvæð, eða um 66%. Þetta voru orð eins og *falleg* ‘beautiful’, sæt ‘cute’ og *svakalega fín* ‘gorgeous’.[32] Aftur á móti höfðu

[32] Hér skal tekið fram að íslensku þýðingarnar koma frá Google Translate en eru ekki þýðingar höfunda. Því getur komið fram misræmi á orðflokkum íslensku þýðinganna.

aðeins 23% orðanna sem fengu karlkynsbeygingu jákvæða merkingu, eða 7 orð af 31. Þetta voru lýsingarorð eins og *myndarlegur* 'handsome', vöðvastæltur 'muscular' og *kynþokkafullur* 'sexy'.

Ólíkt lýsingarorðum sem lýsa persónueinkennum fólks virðast jákvæð lýsingarorð sem lýsa útliti fremur fá kvenkynsbeygingu en karlkynsbeygingu. Þau orð sem lýsa útliti á neikvæðan hátt virðast jafnframt líklegri til þess að fá karlkynsbeygingu en kvenkynsbeygingu. Þetta má sjá á *mynd 3* þar sem tæp 50% þeirra orða sem birtust í karlkyni voru neikvæð. Þetta voru lýsingarorð eins og t.d. *ljótur* 'ugly', *feitur* 'fat' og *viðbjóðslegur* 'repugnant'. Aðeins 33% þeirra orða sem fengu kvenkyn voru neikvæð og fólu flest í sér stærð eða þyngd samanber *of þung* 'overweight', þung 'heavy', *mikil* 'colossal'.

5.3 Gerendur í atburðum

Eins og greint var frá í 4. kafla voru einnig skoðuð orð sem fólu í sér hæfni í ýmsum störfum. Það var gert með því að slá inn enskar setningar á borð við *I am good at sewing/welding* 'ég er góð/ur í að sauma/sjóða' þar sem orð fyrir mismunandi störf komu í stað undirstrikuðu orðanna. Þá var sjónum helst beint að störfum sem tengdust heimilishaldi annars vegar og iðnaðartengdum störfum hins vegar. Í íslenskri þýðingu verður lýsingarorðið *góð/ur* 'good' að hafa málfræðilegt kyn og í þessum flokki virtust það vera tegundir starfanna sem höfðu áhrif á það hvaða kynbeygingu orðið fékk.

Þegar orð tengd iðnaðarstörfum voru prófuð í þessu samhengi kom í ljós að lýsingarorðið fékk karlkyn í 12 af 15 dæmum. Þetta voru orð eins og *trésmíði* 'woodworking', *steypa* 'mortaring' og *framkvæmdir* 'construction'. Í þremur tilvikum birtist setningin í kvenkyni, með orðunum *keyra*, *bora* og *smíða*. Það sem kom á óvart var að þegar þýðingin birtist í kvenkyni varð hún ekki ég er ***góð*** í að ___ eins og við var að búast, heldur varð hún ég er ***dugleg*** *að* ___. Dæmi um þetta má sjá á *mynd 4*.

I am good at driving	Ég er **dugleg** að keyra	kvk
I am good at driving trucks	Ég er **góður** í að keyra vörubíla	kk

Mynd 4. Dæmi um mun á þýðingum milli karlkyns og kvenkyns

Í dæminu er það einnig áhugavert að það eina sem var öðruvísi í inntakssetningunni var að orðinu *trucks* var bætt við. Það leiddi til þess lýsingarorðið fékk ekki aðeins mismunandi þýðingu eftir kyni heldur kom líka fram merkingarlegur munur. Sú staðreynd að setningin birtist í karlkyni þegar

orðinu *vörubíll* var bætt við bendir til að gögnin endurspegli að það sé frekar talið karlmannsstarf að keyra vörubíla.

Í framhaldi af þessu var ákveðið að prófa að bæta neituninni *not* við inntakssetninguna, sbr. *I am* ***not*** *good at* ___ 'ég er **ekki** góð/ur í að ___'. Þá var þýðingin undantekningarlaust í karlkyni, sbr. ég er ekki ***góður*** *(kk) í að* ___. Það bendir til að það sé eitthvað í þýðingarlíkaninu sem gerir það að verkum að setningar með neitun þýðast með þessum hætti.

Þegar orð tengd heimilisstörfum voru könnuð kom fram önnur mynd en af iðnaðarstörfunum. Þá var mun líklegra að lýsingarorðið kæmi út í kvenkyni en karlkyni í íslenskri þýðingu. Þetta voru orð eins og þrífa, þvo, elda, prjóna og baka. Af 21 setningu með mismunandi heimilisstörfum komu 18 setningar út í kvenkyni, eða rúm 85%. Aðeins þrjár setningar fengu karlkyn en þær innihéldu orðin *garðrækt* 'gardening', *ryksuga* 'vacuuming' og *grilla* 'barbequing'. Það er því greinilegt að hugbúnaðurinn reiknar út að það sé tölfræðilega líklegast að konur sinni frekar heimilisstörfum en karlar. Þegar neitun var bætt við þessar setningar kom það sama í ljós og áður að þýðingin var alltaf í karlkyni.

Rétt eins og sást í þau skipti þar sem kvenkynsbeyging kom fram með iðnaðarstörfum varð þýðingin í kvenkyni með heimilisstörfum undantekningalaust ég er ***dugleg*** *að* ___ en ekki ég er ***góð*** í að ___. Lýsingarorðið *good* fékk því aldrei þýðinguna *góð* í kvenkyni, aðeins *dugleg*. Þetta gefur til kynna að umræðan um konur sé á þá leið að þær séu líklegri til að vera *duglegar* að gera eitthvað, fremur en að þær séu raunverulega *góðar* í því.

Af þremur skiptum sem setning um heimilisstörf fékk þýðingu í karlkyni var lýsingarorðið aðeins einu sinni þýtt sem *duglegur*. Það var með orðinu *að ryksuga* 'vacuuming'. Í hinum tveimur fékk það þýðinguna *góður* eins og algengara var með birtingu karlkynsbeygingar. Þennan ótvíræða mun á þýðingum má sjá á *mynd 5* þar sem iðnaðar- og heimilisstörfum er stillt upp hlið við hlið.

Iðnaðarstörf		**Heimilisstörf**	
I am good at electrical work	Ég er **góður** í rafmagnsvinnu	I am good at cooking	Ég er **dugleg** að elda
I am good at mortaring	Ég er **góður** í að steypa	I am good at knitting	Ég er **dugleg** að prjóna
I am good at woodworking	Ég er **góður** í trésmíði	I am good at sewing	Ég er **dugleg** að sauma
I am good at construction	Ég er **góður** í framkvæmdum	I am good at cleaning	Ég er **dugleg** að þrífa

Mynd 5. Dæmi um mun á þýðingum milli starfa

Myndin sýnir að munurinn milli kynja í þýðingunni var ekki aðeins beygingarlegur heldur einnig merkingarlegur. Þetta er áhugavert með samfélags-

legar hugmyndir um kynjahlutverk í huga. Nánar verður rætt um þetta í 6. kafla.

6. Umræður

Í undanfarandi köflum hefur verið fjallað um tæknihalla í máltækni, vandamálin sem hann getur haft í för með sér og mikilvægi þess að benda á hann. Í þeim tilgangi var greint frá niðurstöðum rannsóknar á íslenskum vélþýðingum úr þýðingarforritinu Google Translate. Líkt og fram hefur komið birtist mikill meirihluti þýðinganna í karlkyni sem þarf ekki að koma á óvart því að karlkyn er ómarkað kyn í íslensku. Aftur á móti er áhugavert að skoða þau orð sem birtast í kvenkyni þar sem afgerandi meirihluti þeirra fellur undir sams konar mynstur. Í flokki orða sem lýsa persónuleika og persónueinkennum fólks hefur meirihluti þeirra sem fær kvenkynsbeygingu neikvæða merkingu, þ.e. þau lýsa eiginleikum sem yfirleitt eru taldir neikvæðir. Sem dæmi um þetta má nefna orð eins og *geðveik*, *þunglynd*, *stressuð*, *pirruð*, *viðkvæm*, *óörugg* og fleiri. Orð sem hafa jákvæða merkingu eru aftur á móti mun líklegri til að fá karlkynsbeygingu en það á til dæmis við um orð eins og *sterkur*, *snjall*, *menntaður*, *hugrakkur* og *sjálfsöruggur*.

Þegar kemur að orðum sem lýsa útliti er líklegra að þau fái kvenkynsbeygingu ef þau hafa jákvæða merkingu og er þá átt við orð eins og *falleg* og *sæt*. Þessar niðurstöður eru áhugaverðar í ljósi þess að ákveðin tilhneiging virðist ríkja í samfélaginu til að leggja áherslu á útlit kvenna fremur en aðra jákvæða eiginleika í fari þeirra. Sem dæmi má nefna fréttaflutning um íslenskar vísindakonur sem hafa verið í kastljósi fjölmiðla starfs síns vegna þar sem útlit þeirra verður að aðalumfjöllunarefninu frekar en mikilvæg störf þeirra innan vísindanna. Þarna má til dæmis nefna umræðu um klæðnað Ölmu Möller landlæknis[33] og gleraugu Kristínar Jónsdóttur eldfjalla- og jarðskjálftafræðings.[34] Neikvæð lýsingarorð sem fengu kvenkynsbeygingu voru heldur færri en áhugavert er að þau fólu flest í sér stærð eða þyngd, líkt og fram hefur komið. Aftur á móti var erfitt að greina mynstur í merkingu þeirra orða sem fengu karlkynsbeygingu.

Þá var einnig fjallað um birtingarmynd lýsingarorða sem fela í sér hæfni í

[33] Smartland Mörtu Maríu, „Steldu stíl Ölmu Möller landlæknis“, *Mbl.is*, 2020, sótt 12. apríl 2021 af https://www.mbl.is/smartland/tiska/2020/03/17/steldu_stil_olmu_moller_landlaeknis/.

[34] Smartland Mörtu Maríu, „Gleraugu Kristínar gjörbreyttu útlitinu“, *Mbl.is*, 2021, sótt 12. apríl 2021 af https://www.mbl.is/smartland/tiska/2021/03/04/gleraugu_kristinar_gjorbreyttu_utlitinu/.

tilteknum störfum. Áhersla var lögð á heimilisstörf annars vegar og iðnaðarstörf hins vegar. Niðurstöður sýna greinilegan kynjahalla þar sem orð tengd góðri færni í heimilisstörfum fengu nær alltaf kvenkynsbeygingu, til dæmis *baka*, þrífa, þvo, *elda*, *prjóna*. Þegar neitun var bætt við setninguna breyttust orðin jafnframt á þann hátt að þau fengu karlkynsbeygingu í öllum tilvikum. Þá fengu orð tengd góðri færni í iðnaðarstörfum oftast karlkynsbeygingu, til dæmis *laga* og *steypa*, og þegar neitun var bætt við hélst sú beyging. Kvenkyn var því varla sýnilegt þegar kom að iðnaðarstörfum.

Jafnframt kom fram áhugaverður munur á þýðingum eftir því hvort þau birtust í karl- eða kvenkyni. Þegar þýðingin var í kvenkyni hafði það áhrif á alla setninguna þar sem merkingin breyttist frá því að vera góð/ur í einhverju yfir í að vera dugleg/ur í einhverju. Þannig verður setning á borð við *I am good at cooking* að Ég er dugleg að elda en setning á borð við *I am good at woodworking* að Ég er góður í trésmíði. Erfitt er að skýra hvers vegna þetta gerist en það er greinilegur merkingarlegur munur á því að vera góð/ur í einhverju og dugleg/ur í einhverju. Þannig kemur til dæmis fram nokkuð skýr blæbrigðamunur í því að segja að einhver sé góð/ur í að elda og að einhver sé dugleg/ur að elda. Blæbrigðamunurinn stafar líklegast af því að það að vera dugleg/ur felur ekki í sér vísun í færni og/eða hæfni að sama marki og það að vera góð/ur í einhverju. Það er því hægt að vera dugleg/ur í einhverju án þess að vera sérstaklega góð/ur í því.

Þessar niðurstöður eru að vissu leyti sláandi þar sem kynjahallinn er verulega áberandi en um leið fullkomlega ósjálfráður. Hann er einfaldlega afleiðing þess að þýðingarkerfið er þjálfað á textum frá raunverulegu fólki og því hljóta orðin sem um ræðir að koma oftar fyrir með þessum hætti í gögnunum. Því reiknar forritið út að það sé tölfræðilega líklegast að orðin eigi að birtast svona. Eins og titill greinarinnar gefur til kynna er ekki hægt að segja að þýðingar sem innihalda slíkan réttlætisbrest séu sérlega „góðar“ og það er því með ráði gert að tala um *vondar* vélþýðingar í stað *lélegra*. Ástæðan er sú að ekki er beinlínis um lélegar þýðingar að ræða þar sem orðin eru rétt þýdd merkingarlega séð. Hins vegar má greinilega sjá ákveðin mynstur koma fram í þýðingunum hvað varðar kyn og kynjahlutverk, líkt og fjallað hefur verið um. Þess vegna tölum við um að þýðingarnar séu *vondar* þar sem þær geta ýtt undir ójöfnuð sem er nú þegar til staðar í samfélaginu. Ójöfnuðurinn kemur fram vegna þess að Google Translate byggir á raunverulegum málgögnum og því eru niðurstöðurnar sem sýndar eru í 5. kafla dæmi um ósjálfráða hlutdrægni þar sem þær endurspegla í raun aðeins samfélagslegan halla sem er þegar til staðar.

Eins og kom fram í 2. kafla skiptir máli að vera meðvituð um hvaðan gögnin koma sem mállíkön og þýðingakerfi eru þjálfuð á, hver það eru sem skrifa textana og hvernig orðræða fer þar fram. Ekki er nóg að benda aðeins á vandamálin heldur þarf líka að finna lausnir á þeim. Ein leið til þess að koma í veg fyrir að hlutdrægni komi fram í málgögnum er að hreinsa burt texta sem innihalda ákveðin „tabú-orð" sem eru líkleg til að koma fram í hatursorðræðu og öðrum textum sem geta verið særandi, þá sérstaklega gagnvart ákveðnum hópum samfélagsins.[35] Það getur aftur á móti haft þveröfugar afleiðingar eins og kemur fram í grein Bender o.fl.[36] Með því að útiloka alfarið texta sem innihalda þessi orð er ekki aðeins verið að koma í veg fyrir hatursorðræðu heldur einnig verið að útiloka að jákvæð og þörf umræða um þessi málefni verði hluti af gögnunum. Það verður til þess að textar sem skrifaðir eru um og af fólki sem tilheyrir jaðarsettum hópum eru í miklum minnihluta og þjálfunargögnin endurspegla þá miklu fremur sjónarhorn fólks í forréttindastöðu.

Annað sem bent hefur verið á að geti verið gagnlegt við þróun mállíkana er að endurhugsa uppbyggingu þjálfunargagnanna. Þannig gæti það reynst betur að einblína á uppruna og gæði textanna frekar en einungis gagnamagnið sjálft. Mögulega á hugmyndin „því stærra, því betra" ekki eins vel við málheildir til þjálfunar mállíkana og áður var talið. Í þessu samhengi má nefna meistararitgerð Tinnu Þuríðar Sigurðardóttur í máltækni, en þar metur hún íslenskar málheildir á borð við Risamálheildina og bendir á skekkjur tengdar gagnasöfnun og félagslegum þáttum, þar á meðal kynjahalla. Sem dæmi nefnir hún að Risamálheildin, sem inniheldur 1,5 milljarð orða, samanstandi að mestu leyti af fréttatextum og alþingisræðum, en í því umhverfi hefur mikið hallað á konur í gegnum tíðina. Þess vegna er líklegt að kynjahalli birtist í henni og líkönum sem byggja á henni. Tinna skoðar sérstaklega þær íslensku málheildir sem mest hafa verið notaðar í rannsóknum og þróun máltæknilausna og sýnir m.a. að mikill meirihluti nafngreindra textahöfunda í þessum gagnasöfnum er karlkyns. Tinna sýnir enn fremur fram á að með tiltölulega einföldum lausnum sé hægt að auka fjölbreytileika stærri mál-

[35] Dæmi um slíkan lista fyrir íslensku, sem má nota í margvíslegum tilgangi, er lýst í Agnes Sólmundsdóttir, Lilja Björk Stefánsdóttir og Anton Karl Ingason. „IceTaboo: A database of contextually inappropriate words for Icelandic", *CLARIN Annual Conference 2021 Proceedings*, ritstj. Monica Monachini og Maria Eskevich, bls. 39–42. Sótt 23. ágúst 2021 af https://office.clarin.eu/v/CE-2021-1923-CLARIN2021_ConferenceProceedings.pdf.

[36] Emily Bender o.fl., „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", bls. 614.

heilda. Hún safnaði saman textum úr femíníska tímaritinu *Flóru* og gerði úr þeim málheild sem samanstendur aðeins af um 230 þúsund orðum. Þrátt fyrir smæð hennar komu þar í ljós hundruð orða sem ekki komu fram í Risamálheildinni. Það er merki um að þrátt fyrir að málheildir séu stórar geti textarnir í þeim verið fremur einhæfir.[37]

Eins og greint hefur verið frá er ljóst að það er töluverður munur á birtingarmynd karlkyns annars vegar og kvenkyns hins vegar í þýðingarforritinu Google Translate. Rannsókn okkar sýnir að nokkuð reglubundið mynstur kemur fram í því hvernig orðin þýðast og rök hafa verið færð fyrir því að þetta samsvari að vissu leyti samfélagslegum breytileika í orðræðu í tengslum við mismunandi kyn og hugmyndir fólks um mismunandi kynjahlutverk. Því er fremur ólíklegt að um tilviljun sé að ræða. Ef þau sem þróa búnaðinn eru ómeðvituð um að forritið sýni þennan kynjahalla undirstrikar það jafnframt hversu raunverulegur hann er. Aftur á móti er mögulegt að þau viti af hallanum en telji hann ásættanlegan fórnarkostnað til að ná fram betri þýðingum. Þrátt fyrir að tæknin nái tilsettum árangri getur það þó haft slæmar afleiðingar fyrir samfélagið í heild ef hún leiðir af sér ójöfnuð á borð við kynjahalla.

Google Translate er lokaður hugbúnaður í einkaeigu og upplýsingar um þær málheildir sem forritið byggir á eru ekki aðgengilegar almenningi.[38] Því er erfitt fyrir rannsakendur að átta sig nákvæmlega á hverjir annmarkarnir eru sem valda tæknihalla. Hugbúnaðurinn uppfærist reglulega og því breytast mállíkönin og þar af leiðandi þýðingarnar með tímanum. Þar að auki innleiddi Google árið 2014 þann möguleika að notendur taki virkan þátt í að laga og bæta þýðingarnar, með því að gefa þeim einkunn eða laga þær handvirkt.[39] Enda þótt þetta geri það að verkum að þýðingarvélin þjálfist og verði betri við aukna notkun kemur það ekki í veg fyrir að dulin samfélagsleg mynstur, svo sem kynja- og kynþáttahalli, rati í þýðingarnar. Það stafar af því að notendur sem taka þátt í að bæta þýðingarnar geta ekki verið fullkomlega hlutlausir frekar en þau sem skrifa textana í málheildunum þar sem samfélagsleg áhrif eru ávallt til staðar.

[37] Tinna Þuríður Sigurðardóttir, „When More is Less: Identifying Biases in Large Icelandic Corpora“, MSc-ritgerð í tölvunarfræði við Háskólann í Reykjavík, 2021, http://hdl.handle.net/1946/39430.

[38] Anna Björk Nikulásdóttir o.fl., *Máltækni fyrir íslensku 2018–2022: Verkáætlun*, bls. 75.

[39] Sveta Kelman, „Translate Community: Help us improve Google Translate!“, *Google Inside Search*, 2014, sótt 12. apríl 2021 af https://search.googleblog.com/2014/07/translate-community-help-us-improve.html.

Þar sem hugbúnaðurinn er síbreytilegur á þennan hátt er ekki við því að búast að hægt sé að endurtaka rannsókn okkar og fá sömu niðurstöður, nú meira en ári síðar. Aftur á móti skiptir máli að gera fleiri svona rannsóknir til þess að fylgjast með í hvora áttina þróunin fer svo hægt sé að sporna gegn vandamálum á borð við þessi. Google Translate er í dag mest notaða vélþýðingarkerfið fyrir íslensku og að öllum líkindum það verkfæri sem er mest notað til að búa til tölvugerðan íslenskan texta. Það er því mikilvægt að fylgjast með þessari þýðingarveitu, sérstaklega í ljósi þess að mikil gerjun er í íslenskri máltækni um þessar mundir og vegna opinbers markmiðs stjórnvalda um að gera íslensku gjaldgenga í hinum stafræna heimi.[40]

7. Lokaorð

Markmið þessarar rannsóknar var að kanna hvort kynjahalli kæmi fram í íslenskum þýðingum úr vélþýðingarkerfinu Google Translate, líkt og rannsóknir hafa sýnt fram á að geti gerst í öðrum tungumálum. Rannsóknin leiðir í ljós að sú er raunin og að greinilegt mynstur kemur fram sem samsvarar að vissu leyti samfélagslegum kynjahalla. Líkt og greint var frá kemur mynstrið ósjálfrátt fram vegna þess að forritið byggir á raunverulegum málgögnum og endurspeglar því kynjahalla sem er nú þegar til staðar í samfélaginu. Í nútímasamfélagi hefur orðið mikil vitundarvakning um jafnrétti og hefur í auknum mæli verið litið gagnrýnum augum á söguna hvað það varðar. Kynjahallinn í málgögnunum er þannig hluti af sögunni á þann hátt að skriflegar heimildir varðveita gamlar hugmyndir. Ef ekki er brugðist við gæti ómeðvitaður kynjahalli fortíðarinnar magnað upp kynjamisrétti í framtíðinni. Tæknin skipar sífellt stærri sess í lífi fólks og því er hættulegt að hún vinni gegn þeirri réttlætisbaráttu sem hefur átt sér stað undanfarið. Þessar niðurstöður sýna ótvírætt fram á mikilvægi þess að þróaðar séu máltæknilausnir sem geta borið kennsl á slíkan halla og þar með komið í veg fyrir að hann viðhaldist.

ÚTDRÁTTUR

Greinin fjallar um tæknihalla í máltækni, en það er þegar mállíkön sem þjálfuð eru á raunverulegum málgögnum fara óumbeðin að endurspegla samfélagslegan ójöfnuð á borð við kynjahalla, óháð ásetningi þeirra sem búa kerfin til. Einblínt er á kynja-

[40] Anna Björk Nikulásdóttir o.fl., *Máltækni fyrir íslensku 2018–2022: Verkáætlun*.

halla í vélþýðingum og í kjölfarið er greint frá rannsókn sem gerð var á íslenskum þýðingum Google Translate. Niðurstöður sýna að töluverður munur er á birtingarmynd karlkyns og kvenkyns í þýðingarvélinni. Þar má sjá mynstur sem samsvara ákveðnum samfélagslegum hugmyndum um kyn og kynjahlutverk, en sem dæmi virðist það háð merkingu lýsingarorða sem vísa til fólks hvort þau birtust í karlkyni eða kvenkyni. Þau sem fela í sér jákvæð persónueinkenni birtust frekar í karlkyni en þau neikvæðu í kvenkyni. Þessu var aftur á móti öfugt farið með útlitstengd lýsingarorð. Þar að auki birtust ákveðnar staðalmyndir í þýðingum tengdum konum og heimilisstörfum annars vegar og körlum og iðnaðarstörfum hins vegar. Þessar niðurstöður sýna ótvírætt fram á mikilvægi þess að vera á varðbergi gagnvart tækninni svo hún viðhaldi ekki úreltum samfélagslegum hugmyndum — sér í lagi í hinum stafræna heimi nútímans.

Lykilorð: máltækni, tæknihalli, kynjahalli, vélþýðingar, kyn

ABSTRACT

This paper examines machine bias in language technology. Machine bias can affect machine learning algorithms when language models trained on large corpora include biased human decisions or reflect historical or social inequities, e.g. regarding gender and race. The focus of the paper is on gender bias in machine translation and we discuss a study conducted on Icelandic translations in Google Translate. The results show a pattern which corresponds to certain societal ideas about gender and gender roles. For example it seems to depend on the meaning of adjectives referring to people whether they appear in the masculine or feminine form. Adjectives describing positive personality traits were more likely to appear in masculine gender whereas the negative ones frequently appear in feminine gender. However the opposite applied to appearance related adjectives. Furthermore, certain stereotypes were reflected in the translations regarding women and domestic skills on one hand and men and blue-collar industrial skills on the other. These findings unequivocally demonstrate the importance of being vigilant towards technology so as not to maintain societal inequalities and outdated views — especially in today's digital world.

Key words: language technology, machine bias, gender bias, machine translation (MT), gender

Agnes Sólmundsdóttir
BA-nemi í almennum málvísindum
Íslensku- og menningardeild
Hugvísindasviði Háskóla Íslands
IS-102 Reykjavík, Ísland
ags46@hi.is

Dagbjört Guðmundsdóttir
Doktorsnemi í íslenskri málfræði
Íslensku- og menningardeild
Hugvísindasviði Háskóla Íslands
IS-102 Reykjavík, Ísland
dagu@hi.is

Lilja Björk Stefánsdóttir
Doktorsnemi í íslenskri málfræði
Íslensku- og menningardeild
Hugvísindasviði Háskóla Íslands
IS-102 Reykjavík, Ísland
lbs@hi.is

Anton Karl Ingason
Dósent í íslenskri málfræði og máltækni
Íslensku- og menningardeild
Hugvísindasviði Háskóla Íslands
IS-102 Reykjavík, Ísland
antoni@hi.is