Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3113–3121
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Mean Machine Translations:
On Gender Bias in Icelandic Machine Translations

Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir
Lilja Björk Stefánsdóttir, Anton Karl Ingason
University of Iceland
Sæmundargata 2, 102 Reykjavík, Iceland
{ags46, dagu, lbs, antoni}@hi.is

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 the translation systems Google Translate and Vélþýðing.is. The results show a pattern which corresponds to certain societal ideas about gender. 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. 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demonstrate the importance of being vigilant towards technology so as not to maintain societal inequalities and outdated views — especially in today's digital world.

Keywords: language technology, machine bias, gender bias, machine translation, gender

1. Introduction

The main purpose of this paper is to examine how gender bias appears in English-Icelandic translations. A study was conducted on Icelandic translations in the translation systems Google Translate and Vélþýðing.is.[1] It was conducted over a period of 20 months where Google Translate was tested three times providing data on how the representations of gender bias can fluctuate over time. Vélþýðing.is, a translation tool focusing on Icelandic, was tested alongside the multilingual Google Translate in the latest period to have a comparison between two different translation tools in order to see if a similar gender bias occurs.

Our main research question is: *How does gender-bias appear in English-Icelandic translations and how does it vary over a 20 month period as the systems evolve?*

The paper is structured as follows. Section 2 gives an overview of previous research concerning machine bias, focusing on gender bias in machine translation tools. In Section 3, we describe the methodology of the study and Section 4 gives statistical information on how gender bias is manifested in both translation tools, comparing the results. Finally, we conclude with Section 5.

2. Background

In the field of Machine Translations (MT), statistical and neural network approaches have been dominant in recent years (Zong and Hong, 2018). Statistical and neural machine translation systems are data-driven in the sense that large collections of linguistic data are used to train these models for natural language processing (NLP). Such methods rely heavily on parallel datasets to obtain the linguistic information needed to automatically translate between different languages. These parallel datasets contain large amounts of texts that have been written in one language and manually translated to the other and in order to train the translation models properly, the datasets must be very large (Barkarson and Steingrímsson, 2019).

Linguistic corpora that are used to train language models consist of texts that are written by people. Therefore these texts can contain discourse topics that reflect the views and opinions of said people. In turn, the texts also hold all sorts of societal ideas and patterns created by the culture that has molded the language in some part over time (Prates et al., 2018).

Language models trained on large, uncurated datasets from the Web have been shown to encode hegemonic views that are harmful to marginalized populations (Bender et al., 2021). Up to this point, the main focus in language technology has been on gathering as much data as possible in order to properly train language models and the large size of training data has enabled deep learning models to achieve high accuracy in NLP. However, the training data has been shown to have problematic characteristics, resulting in technology that encodes stereotypical and derogatory associations along the lines of gender, ethnicity, race, disability status and so on. Seeing as language models are so heavily influenced by their training data, it is important to know where the texts come from, e.g. who

[1] The following abbreviations and acronyms are used in the paper: FEM=feminine, MASC=masculine, NLP=Natural Language Processing, STEM=Science, Technology, Engineering and Mathematics.

their writers are in terms of societal factors such as age, class, race, gender and cultural background.

A large amount of data does not necessarily guarantee its diversity (Bender et al., 2021; Sigurðardóttir, 2021). Seeing as the Internet is a large and diverse virtual space, one might assume that large databases of texts collected from the Web should be considered broadly representative of the different views of different people around the world. However that is not the case as there are many factors which narrow Internet participation. For example, a lot of the texts collected in this way come from social media and other open and accessible sites, such as Twitter, Reddit and Wikipedia, and studies have shown that the majority of the users of these media are white, young, Western males (Barera, 2020; Barthel et al., 2016). This indicates that the discourse and topics reflected in the texts are largely derived from very privileged social groups, therefore, a discourse from a hegemonic viewpoint is more likely to be retained rather than other more diverse viewpoints (Bender et al., 2021).

As the neural networks use machine learning to learn languages, i.e. by analyzing these texts and learning their grammatical structure and other patterns of the languages, it is inevitable that they unintentionally obtain additional information from the semantic context on opinions, prejudice and biases, such as gender bias, racial bias, and some other representations of societal ideas that these texts may hold. This can lead to machine translation systems and other data driven language technology solutions reflecting and possibly magnifying certain biases and a variety of negative discourse that is derogatory towards people of certain minorities or other less privileged groups of society (Mehrabi et al., 2021; Bender et al., 2021; Zhao et al., 2018). This is called machine-bias and has been the focus of many recent studies (Kirkpatrick, 2016; Danks and London, 2017; Sólmundsdóttir et al., 2021; Bender et al., 2021).

Machine translations systems such as Google Translate are a good platform for checking whether and how machine-bias such as gender bias appears. Previous studies show that gender bias has been proven to exist in Google Translate in multiple languages (Prates et al., 2018; Schiebinger, 2014b; Schiebinger, 2014a; Vanmassenhove et al., 2018). By translating sentences concerning different professions from gender neutral languages to English, Prates et al. (2019) demonstrated that Google Translate and other current translation tools can exhibit gender biases and a strong preference for male defaults. Not only did their findings show that male defaults were prominent, but they were also exaggerated in professions associated with gender stereotypes, such as STEM (Science, Technology, Engineering and Mathematics) occupations. This is in accordance to reigning societal ideas concerning the genders.

Gender stereotypes are broad generalizations about what men and women are like, and they are usually widely accepted (Hentschel et al., 2019). An example of different stereotypes of the genders is the idea that women are valued more for their appearance than men. Women are strongly urged to be physically attractive by the society, i.e. media promotes attention to their appearance in such a way that suggests that looks are their most valued attributes. However, men seem to be valued more for their strong personal traits, such as regarding intelligence, decisiveness and leadership, while women can be criticized for the same traits (Rudman et al., 2012; Rudman and Glick, 2021). These stereotypes along with many other ideas concerning the so-called innate difference between men and women are one of the ways in which gender bias appears in the output of machine translation systems such as Google Translate. This is further amplified in translation outputs of gendered languages where the input language is gender neutral such as English-Icelandic as will be described in section 3.

Icelandic has three grammatical genders: masculine, feminine and neuter. Nouns have fixed gender whereas adjectives receive their gender from the nouns they describe or refer to. The masculine gender in Icelandic is the default gender in many situations, meaning it has the widest range of application whereas the feminine gender has the smallest range. English adjectives and nouns on the other hand do not differentiate between genders, in fact one of the only remaining forms of grammatical gender in English is found in third person pronouns *he/she* (Rögnvaldsson, 1990; Kvaran, 2005).

3. Methodology

Google Translate is one of the most prominent machine translation tools that are publicly available, amounting to 500 million users (Turovsky, 2016). For Icelandic, Google Translate is without a doubt the most used translation tool. The most recent translation system for Icelandic is, however, Vélþýðing.is, currently being developed by the company Miðeind. The system follows tried and tested methods in neural machine translation and is a part of the government sponsored Icelandic Language Technology Programme (Nikulásdóttir et al., 2020). The main emphasis is to build an accessible and effective translation system for Icelandic (Símonarson et al., 2021).

In this study we assume that gender bias in machine translations can be assessed by translating gender neutral sentences in English to Icelandic, where grammatical gender is necessary, by means of an automated translation tool. In this context, we used the two translation tools, Google Translate and Vélþýðing.is, to translate English sentences, containing a first person pronoun and an adjective, to Icelandic. When first person pronouns in English are used to refer to a person, there is nothing in the structure of the sentence that specifies that person's natural gender. However, the adjectives in Icelandic always show the gender of the person being referred to. Therefore, the translation models have to calculate which gender should appear when translating to Icelandic. While it would be interesting to study the internal details of the translation systems, our study focuses on the output of the system from the perspective of the user.

We compiled a list of adjectives (N=321)[2] that are generally used to describe people and classified them by meaning, i.e. if they are used to describe people in a negative, positive or neutral way. The adjectives were collected manually through brainstorming sessions at the Language and Technology Lab. A similar brainstorming method has been previously applied at the lab to compile word lists with good results (Sólmundsdóttir et al., 2021) in cases where a suitable language resource was not previously available. In addition, these adjectives were classified into two categories, based on our expectations derived from well known cultural stereotypes; adjectives that describe personality traits (N=256), e.g. *strong, weak, clever* and *stupid*, and adjectives that describe appearance (N=65), e.g. *beautiful, ugly, fat* and *thin*. These words can be found in Tabels 1 and 2. The output was then classified by the gender of the adjectives in the Icelandic translations and analyzed to see if the translations show a gender bias. The structure of the sentences can be seen in (1) and (2):

(1) Ég er **sterkur**
I am strong.MASC

(2) Ég er **sterk**
I am strong.FEM

Each sentence starts with a 1st person singular pronoun, the verb 'to be', as well an adjective that agrees with the semantic gender of the discourse referent picked out by the pronoun. We only focused on the 1st person singular. It would be interesting to apply this kind of a methodology to the 2nd person as well as the plural but we leave such inquiries for future work.

This study was conducted over a period of 20 months. We ran the sentences through Google Translate three times, in March 2020, March 2021 and October 2021, to be able to compare the results and see if there were any changes regarding gender bias in the translation tool. Vélþýðing.is was used alongside Google Translate at the latest time period, in October 2021, to have comparison between two different translation tools, with one being especially developed for Icelandic.

4. Results

Figure 1 shows results from Google Translate on sentences containing adjectives describing personality traits. It is evident that words describing positive personality traits, such as *strong* and *clever*, appear more often in the masculine form rather than feminine, whereas negative ones, such as *weak* and *stupid* are more likely to appear in the feminine form. In March 2020 the difference between genders is most exaggerated as 60% of the words which appear in masculine form are positive and 60% of those that appear in feminine are negative. Despite these numbers fluctuating somewhat over the 20 month period, the pattern is always the same. Sentences that describe different people's personality traits in a positive way appear most often in masculine form while the majority of the negative ones appear in feminine.

Figure 2 shows the comparison of results on adjectives describing different personality traits between Vélþýðing.is and Google Translate in October 2021. Contrary to Figure 1 the results of Vélþýðing.is do not mirror the same pattern. The ratio of positive, negative and neutral words is relatively the same for words appearing in both masculine and feminine gender. This suggests that gender bias is not as prominent in the output of Vélþýðing.is as it is in Google Translate.

Figure 3 and 4 show results for sentences containing adjectives which describe peoples appearance. Figure 3 demonstrates how these words appear in masculine or feminine form in Google Translate over the 20 month period. The patterns which appear are in some ways opposite to the one in Figures 1 and 2, i.e. words describing peoples appearances negatively, such as *ugly* and *nasty*, are most likely to appear in the masculine form, whereas words describing appearance in a positive way, such as *pretty* and *gorgeous*, appear most often in the feminine. The difference is most noticeable in March 2020, just as in Figure 1, however the pattern almost disappears in March 2021 meaning that the curve of the ratio between positive, negative and neutral words is mostly flattened. This might be considered as a step in the direction of erasing gender bias from the system but the fact that the pattern appears again in October 2021, although not as extreme, suggests that bias is still present.

Figure 4 demonstrates how the results of this category from Vélþýðing.is compare to those of Google Translate in October 2021. Overlooking that the majority of words in the masculine form are neutral, the pattern of positive and negative words describing appearance is relatively the same as in Google Translate, meaning there are more negative words in the masculine than there are positive, although the difference is quite small. It should be noted that in the category of appearance-related words, there does not seem to be as much of a difference as in the category of personality traits, however, this category included fewer words, which could skew the results.

These results show a pattern which corresponds to certain societal ideas about gender. For example it seems to depend on the meaning of adjectives referring to people whether they appear in the masculine or fem-

[2] Note that originally, 446 words were tested. However, 103 words were filtered out due to either wrong translations, translations where words appear the same in masculine and feminine form or translations where the syntax changed so that grammatical gender was not detectable.

inine 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, as words describing what is generally thought to be desirable appearance factors appeared in the feminine gender whereas negative adjectives appeared in the masculine.

5. Conclusion

The main purpose of this paper was to examine how gender-bias appears in English-Icelandic translations. A study was conducted on Icelandic translations in the translation systems Google Translate and Vélþýðing.is, with the following research question as an aim: *How does gender-bias appear in English-Icelandic translations and how does it vary over a 20 month period as the systems evolve?*

The results show that there seems to be a correspondence between the grammatical genders in which the translations appear and certain societal ideas about the genders. This can therefore be interpreted as a direct consequence of the representations of gender in the training data. Seeing as discourse on women is mostly concerned with looks and their appearance is considered their most valued feature, it is of no surprise that adjectives conveying meaning of positive appearance traits are more likely to appear in the feminine form. Likewise, the dominant discourse on men, for example their good skills in leadership and intelligence, is reflected in the results of words describing positive personality traits appearing most often in the masculine form. This is consistent to the gender stereotypes discussed in Section 2.

The results on how gender bias appears in Google Translate over a 20 month period shows that while the numbers fluctuate somewhat, from extreme bias to a relatively small difference between genders, the patterns of gender bias still exists in the translation system. Despite the results demonstrating that gender bias exists in both translation systems, it seems that it is not as prominent in the output of Vélþýðing.is as in Google Translate. This is interesting bearing in mind that the latter is a widely used multilingual translation tool created by a tech company which dominates its field, while Vélþýðing.is is specifically created for the Icelandic language and its small language community.

There is by now a considerable awareness of equality in modern society and social changes are constantly moving in that direction. People have thus become increasingly critical of deep-rooted notions of gender and gender roles in history. Written sources, on the other hand, preserve the ideas of their time, so language data collected from these sources can not be considered fully descriptive of modern society. If left unmanaged, we risk that biases of the past will amplify discrimination in the future. It is important to continually monitor new technology and prevent it from perpetuating outdated societal ideas such as gender bias, because it is imperative that advances in technology must not hinder positive social change - especially in today's digital world.

6. Bibliographical References

Barera, M. (2020). Mind the gap: Addressing structural equity and inclusion on Wikipedia.

Barkarson, S. and Steingrímsson, S. (2019). Compiling and filtering parice: An English-Icelandic parallel corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145.

Barthel, M., Stocking, G., Holcomb, J., and Mitchell, A. (2016). Seven-in-Ten Reddit users get news on the site. Pew Research Center's Journalism Project.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Danks, D. and London, A. J. (2017). Algorithmic bias in autonomous systems. In *IJCAI*, volume 17, pages 4691–4697.

Hentschel, T., Heilman, M. E., and Peus, C. V. (2019). The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in psychology*, 10:11.

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Communications of the ACM*, 59(10):16–17.

Kvaran, G. (2005). *Íslenskt tunga II. Orð.* Almenna bókafélagið.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul.

Nikulásdóttir, A., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422.

Prates, M. O., Avelar, P. H., and Lamb, L. (2018). Assessing gender bias in machine translation–a case study with Google translate. *arXiv preprint arXiv:1809.02208.*

Rudman, L. A. and Glick, P. (2021). *The social psychology of gender: How power and intimacy shape gender relations. Second edition.* Guilford Publications.

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., and Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology*, 48(1):165–179.

Rögnvaldsson, E. (1990). *Íslensk orðhlutafræði.*

Kennslukver handa nemendum á háskólastigi. Fourth edition. Málvísindastofnun Háskóla Íslands.

Schiebinger, L. (2014a). Gendered innovations: Harnessing the creative power of sex and gender analysis to discover new ideas and develop new technologies. *Triple Helix*, 1(1):1–17.

Schiebinger, L. (2014b). Scientific research must take gender into account. *Nature News*, 507(7490):9.

Sigurðardóttir, T. Þ. (2021). When more is less: identifying biases in large Icelandic corpora. Master's thesis, Reykjavík University.

Símonarson, H. B., Snæbjarnarson, V., Ragnarson, P. O., Jónsson, H., and Þorsteinsson, V. (2021). Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139.

Sólmundsdóttir, A., Stefánsdóttir, L. B., and Ingason, A. K. (2021). Icetaboo: A database of contextually inappropriate words for Icelandic. In Monica Monachini et al., editors, *Proceedings of CLARIN Annual Conference 2021*, pages 39–43.

Sólmundsdóttir, A., Guðmundsdóttir, D., Stefánsdóttir, L. B., and Ingason, A. K. (2021). Vondar vélþýðingar: Um kynjahalla í íslenskum þýðingum google translate. *Ritið*, 21(3):177–200.

Turovsky, B. (2016). Ten years of Google Translate. *Google Official Blog*.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zong, Z. and Hong, C. (2018). On application of natural language processing in machine translation. In *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pages 506–510. IEEE.

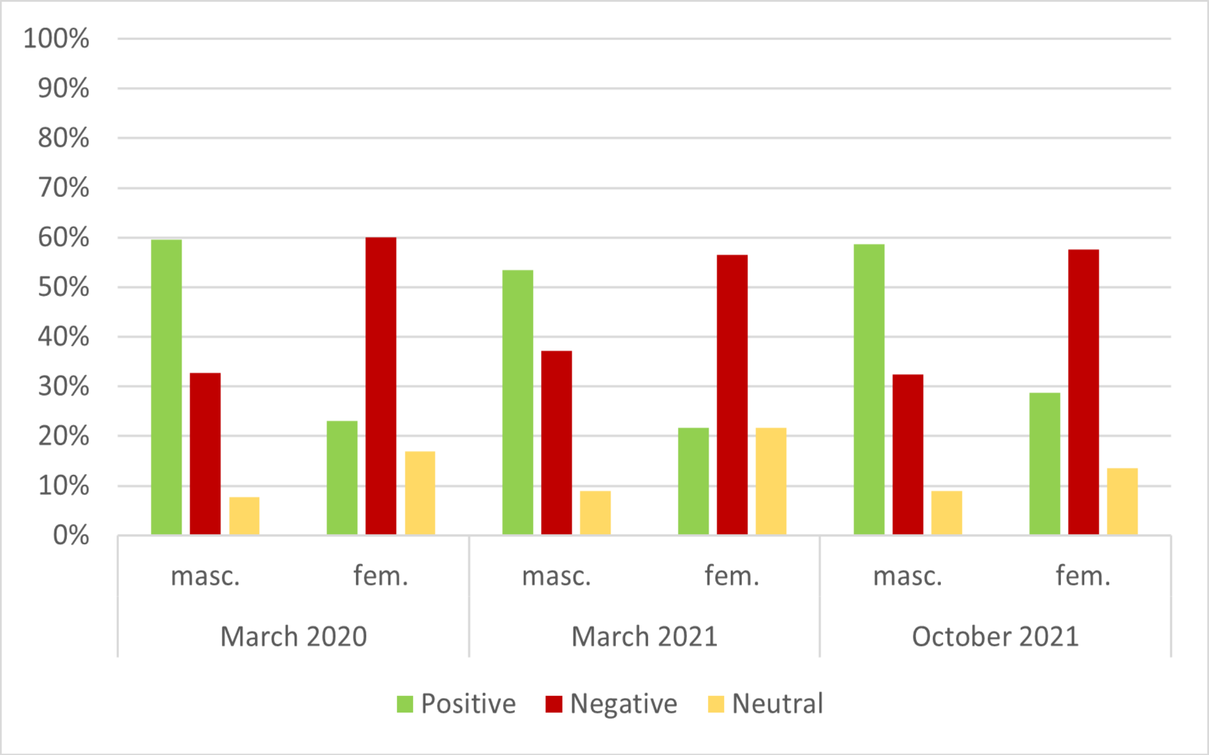

Figure 1: Ratio of positive, negative and neutral adjectives describing personality traits in Google Translate by gender over a 20 month period.

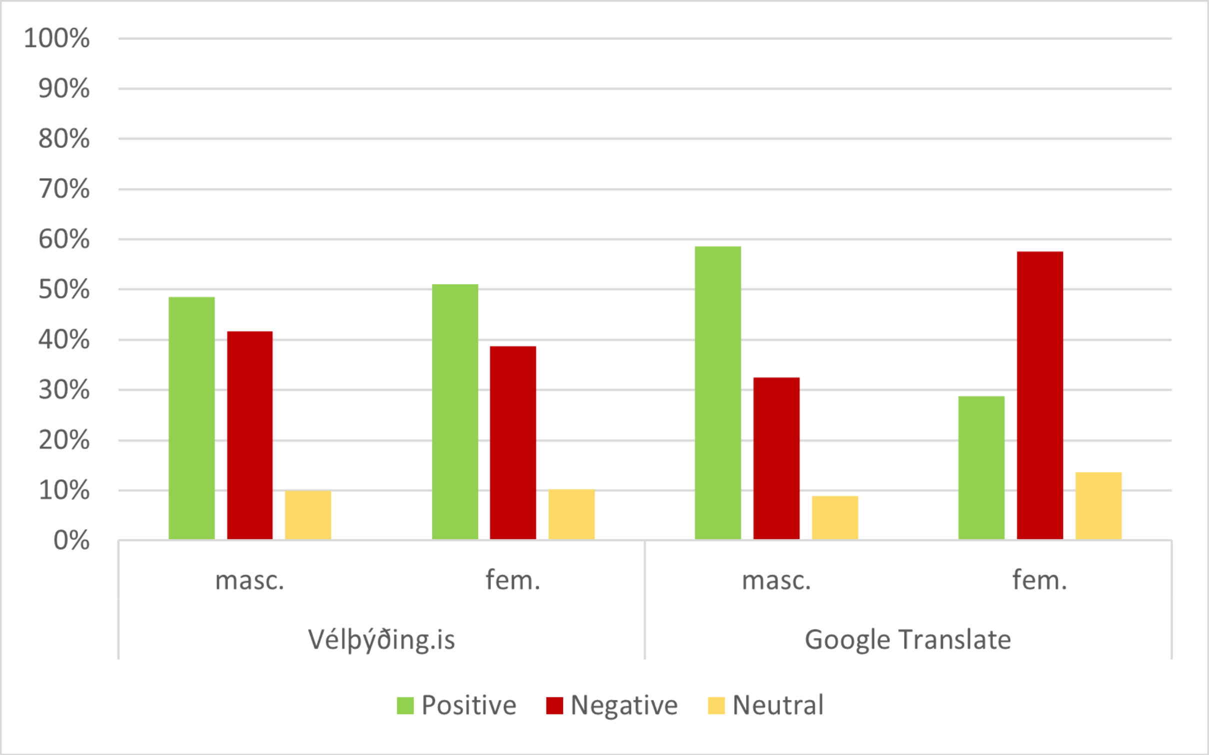

Figure 2: Ratio of positive, negative and neutral adjectives describing personality traits in Vélþýðing.is and Google Translate by gender in October 2021.

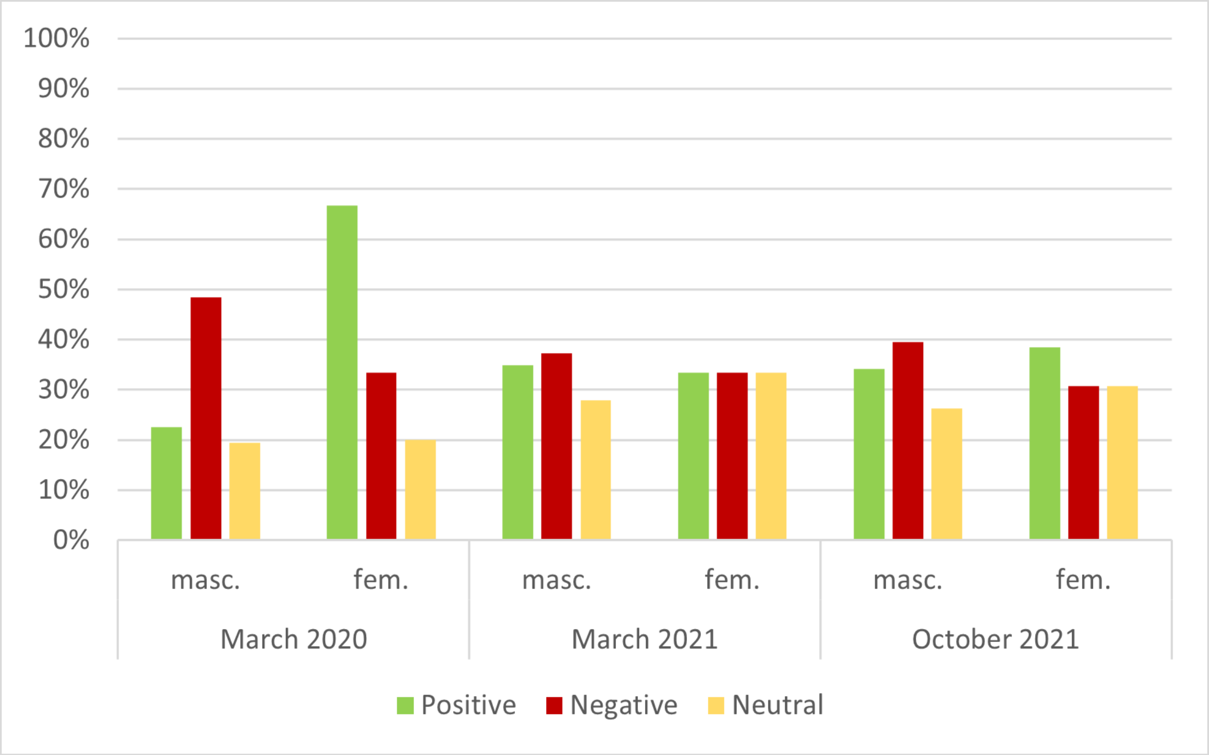

Figure 3: Ratio of positive, negative and neutral adjectives describing appearance in Google Translate by gender over a 20 month period.

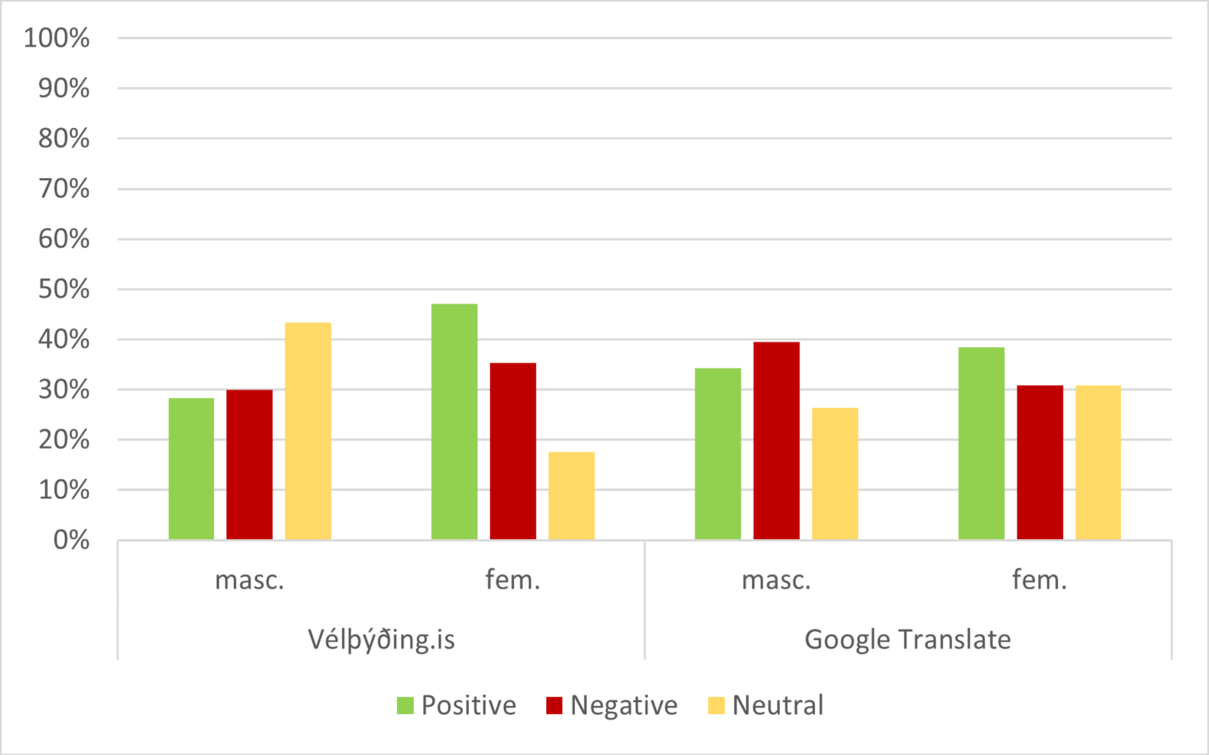

Figure 4: Ratio of positive, negative and neutral adjectives describing appearance in Vélþýðing.is and Google Translate by gender in October 2021.

Adjectives that describe personality traits: *"I am..."*					
absurd	condemned	envious	impressive	odd	stout
accomplished	confident	evil	inatriculite	old	strong
admirable	confused	excited	incapable	oldish	stupid
adorable	considerate	fair	inconsiderate	ordinary	superficial
affable	cool	faithful	inconvenient	passionate	sweet
aggressive	courageous	famous	indelicate	patient	talented
alert	cowardly	fancy	influential	perfect	tame
amazing	crazy	fantastic	insane	pitiful	tender
ambitious	creepy	feeble	insecure	pleasant	terrible
amiable	critical	fierce	insignificant	polite	thankful
angry	crude	filthy	intelligent	poor	the hottest
annoyed	cruel	fine	interesting	popular	thoughtless
anxious	dainty	firm	irrational	positive	tidy
arrogant	dangerous	focused	jealous	powerful	tough
articulate	dark	fracturable	jolly	professional	triumphant
arty	dead	fragile	judicious	proud	unclean
awesome	deceased	frail	kind	psycho	undiplomatic
awful	defiant	friable	layered	psychotic	unfaithful
awkward	delicate	friendly	lazy	quiet	unimportant
bad	delightful	frightened	lesbian	rare	unintelligent
bewildered	depressed	fun	lethargic	rational	uninteresting
bisexual	determined	funny	lively	ready	unlucky
bizarre	different	gay	lost	reasonable	unpleasing
bloody	difficult	generous	lovely	repulsive	unpopular
boring	diplomatic	gentle	lucky	rich	unprofessional
brave	dirty	gifted	mad	robust	unusual
brawny	dishonest	glorious	magnificent	rude	uptight
breakable	divine	good	marvelous	rugged	vicious
bright	doubtful	graceful	mellow	sane	victorious
brittle	dull	great	mighty	scared	vulnerable
busy	dumb	gregarious	mild	scary	wayward
calm	durable	grotesque	modern	selfish	weak
careful	dynamic	hard-working	modest	sensible	weird
cautious	eager	healthy	moody	shallow	wild
charming	educated	helpful	mysterious	shatterable	wise
cheerful	elderly	helpless	negative	shrill	witty
classy	elegant	heterosexual	nervous	shy	wonderful
clean	eloquent	honest	neurotic	significant	worried
clever	embarrassing	horrible	nice	silly	young
clumsy	enchanting	hysterical	nonsensical	sincere	youngish
coarse	encouraging	ignorant	normal	smart	zealous
common	energetic	impatient	obedient	soft	
compelling	enthusiastic	imprecise	obnoxious	stalwart	

Table 1: Input sentences concerning personality traits

Adjectives that describe appearance: *"I am..."*

attractive
bald
beautiful
big
black
blonde
blue
broad
bumpy
chubby
colossal
curvy
cute
elderly
fat
feminine
fit
gigantic
gorgeous
grand
greasy
hairy
handsome
heavy
huge
immense
lanky
large
little
loose
masculine
muscular
nasty
neat
obese
overweight
painted
plain
polished
presentable
pretty
redheaded
repugnant
rough
sexy
shapely
short
skinny
slender
slim
small
spotted
stooped
stunning
symmetrical
tall
tan
thick
thin
tiny
tough
tremendous
ugly
well

Table 2: Input sentences concerning appearance