How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns

Anonymous ACL submission

Abstract

Gender-neutral pronouns have recently been introduced in many languages to a) include non-binary people and b) as a generic singular. Recent results from psycho-linguistics suggest that gender-neutral pronouns (in Swedish) are *not* associated with human processing difficulties. This, we show, is in sharp contrast with automated processing. We show that gender-neutral pronouns in Danish, English, and Swedish are associated with higher perplexity, more dispersed attention patterns, and worse downstream performance. We argue that such conservativity in language models may limit widespread adoption of gender-neutral pronouns and must therefore be resolved.

1 Introduction

Many linguistic scholars have observed how technology in general has altered the course of language evolution (Kristiansen et al., 2011; Abbasi, 2020), e.g., through the influence of social media conventions. Language technologies, in particular, have also been argued to have such effects, e.g., by reducing the pressure to acquire multiple languages.

Gender-neutral pronouns is not an entirely modern concept. In 1912, Ella Flag Young, then superintendent of the Chicago public-school system, said the following to a room full of school principals: "The English language is in need of a personal pronoun of the third person, singular number, that will indicate both sexes and will thus eliminate our present awkwardness of speech." The use of gender-neutral pronouns has become much more popular in recent years (Gustafsson Sendén et al., 2021). In 2013, a gender-neutral pronoun was *politically* introduced in Swedish (Gustafsson Sendén et al., 2015) which can be used for both, people identifying outside the gender dichotomy and as a generic pronoun where information about gender is either unavailable or irrelevant.

In a recently recorded eye-tracking study, Vergoossen et al. (2020a) found no evidence that native speakers of Swedish find it harder to process gender-neutral pronouns than gendered pronouns, an argument often brought up by opponents of gender-inclusive language (Speyer and Schleef, 2019; Vergoossen et al., 2020b). In combination with their increasing popularity, this suggests gender-neutral pronouns have been or will be widely and fully adapted over time (Gustafsson Sendén et al., 2015, 2021). However, since language technology has the potential to alter the course of language evolution, we want to make sure that our NLP models do not become a bottleneck for this positive development.

Contribution We extract stimuli from a Swedish eye-tracking study that has shown no increase in processing cost in humans for the gender-neutral pronoun *hen* compared to gendered pronouns. We translate those stimuli into English and Danish and compare model perplexity across gendered and gender-neutral pronouns for all three languages. Furthermore, we systematically investigate performance differences across pronouns in downstream tasks, namely natural language understanding (NLI) and coreference resolution. Across the board, we find that NLP models, unlike humans, are challenged by gender-neutral pronouns, incurring significantly higher losses when gendered pronouns are replaced with their gender-neutral alternatives. We argue this is a problem the NLP community must take seriously.

2 Model perplexity and attention

In this section we introduce a Swedish eye-tracking study and explain how we adapt this study to investigate gender-neutral pronouns in language models.

Humans and *hen* Vergoossen et al. (2020a) recently recorded a Swedish eye-tracking study to test the hypothesis whether the gender neutral pro-

	en			da			sv	
	she/he	they	xe	hun/han	de	høn	hon/han	hen
perplexity	1	1.49	2.37	1	1.21	3.55	1	1.8
correlation	0.12	0.26	0.33	-0.14	0.03	-0.1	0.19	0.09
	0.28	0.33	0.49	0.13	0.16	0.2	0.65	0.72
	0.28	0.33	0.49	0.13	0.17	0.21	0.65	0.72

Table 1: Perplexity scores across pronouns and languages for the eye-tracking stimuli. Correlation between attention flow and perplexity are listed row-wise for layers 1, 6 and 12.

noun *hen* has a higher processing cost during pronoun resolution than gendered pronouns. Participants were reading sentence pairs where the first sentence contained a noun referring to a person and the second sentence contained a pronoun referring to that person either with a gendered pronoun or *hen*. It has recently been shown that attention flow, in contrast to attention itself, correlates with human fixation patterns in task-specific reading (Anonymous, 2022). We applied a similar analysis pipeline here and extracted all 384 sentence pairs and fed them into the uncased Swedish BERT model.[1] We calculate perplexity values for each sentence pair over word probabilities as given by BERT with the formula proposed by Wang et al. (2019). Furthermore, we calculate attention flow propagated from layers 1, 6 and 12 (Abnar and Zuidema, 2020) and extract attention flow values assigned to the pronoun with respect to the entity. Attention flow considers the attention matrices as a graph, where tokens are represented as nodes and attention scores as edges between consecutive layers. The edge values, i.e., attention scores, define the maximal flow possible between a pair of nodes. We consider different parameters of human fixation which we assume might be influenced by a change in pronouns, in particular during pronoun resolution, i.e., first and total fixation time on the pronoun and fixation time after the first fixation on the noun. For both attention flow and perplexity, however we could not find any meaningful correlation to those parameters. One reason for that might be that the dataset only contains fixations for the two entities, i.e., pronoun and noun, which makes data comparably sparse and impossible to extract complete reading patterns.

Language models and gender-neutral pronouns We therefore focus on the model-based data alone in order to understand how well language models can deal with gender-neutral pronouns. For this, we consider perplexity values on sentence-level and calculate rank-based Spearman correlation between perplexity and attention flow for the aforementioned layers. With this analysis, we can see if a) gender-neutral pronouns cause a higher sentence perplexity, i.e., a higher *surprisal* and if b) a possible higher surprisal is connected to higher attention flow values on the pronoun with respect to the entity. We furthermore translate the sentence pairs into English and Danish where we use two sets of gender-neutral pronouns: 3rd person plural (hence: they/de) which are used in both languages as gender-neutral pronouns (Miltersen, 2020) and *neopronouns* (xe for English (Hekanaho, 2020) and høn for Danish).[2] We apply the same experiments to those translated datasets with uncased Danish BERT[3] and uncased English BERT [4].

Results We show results on perplexity and correlations in Table 1 for Danish, English and Swedish. Perplexity values for the datasets with gendered pronouns are set to 1 and we show relative increase for gender-neutral pronouns within a language since perplexity values have been shown to not be comparable across languages (Mielke et al., 2019; Roh et al., 2020). There we can see that perplexity scores for sentences with gender-neutral pronouns are significantly higher (Wilcoxon signed-rank test and received p-values $< .01$ for all pair-wise comparisons). For the correlation between perplexity and attention flow on the Swedish sentence pairs, we can see a clear development between the first layer where there is no correlation ($p > .05$) for gender-neutral *hen* and very low correlation for gendered pronouns which changes for the other layers where correlations for *hen* are even higher

[1] https://huggingface.co/af-ai-center/bert-base-swedish-uncased/tree/main

[2] information.dk/kultur/2020/03/hen-hoen-saadan-kom-nye-pronominer-debatten-sproget

[3] https://huggingface.co/Maltehb/danish-bert-botxo

[4] https://huggingface.co/bert-base-uncased

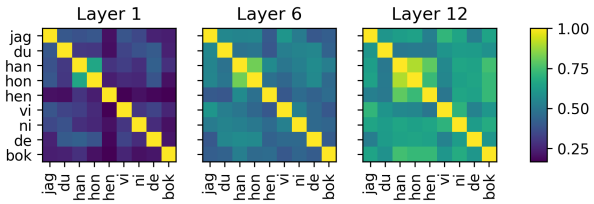

Figure 1: Pair-wise cosine similarity between word representations of all pronouns and the Swedish word *bok* (book) as a baseline for different layers of BERT. We see that gender-neutral *hen* grows from being an outsider (similar to *bok*) in the 1st layer into the cluster of gendered 3rd person pronouns *hon/han* across layers.

($\rho = 0.72$) than for gendered pronouns ($\rho = 0.65$). This suggests that there is some development across layers that is stronger for *hen* than for gendered pronouns. Furthermore, we see a similar evolvement for correlations across layers in English but a much weaker correlation for Danish. To investigate those effects across layers further, we look at word embeddings for all Swedish pronouns from all 12 layers in BERT and compute pair-wise cosine similarity including the Swedish word for book (*bok*) as a baseline where we expect no specific relation to pronouns. In Figure 1, we see less similarity between *hen* and the other pronouns in the first layer. This changes for layer 6 and 12 where word representations seem to be more similar and the three 3rd person pronouns *hen, han, hon* get closer to each other. This is in line with the literature where it has been found that single attention heads perform better on pronoun resolution than others. In particular middle and deeper layers have shown stronger attention weights between coreferential elements (Vaswani et al., 2017; Webster et al., 2018; Clark et al., 2019). Given that we do not consider individual heads or layers but the entire attention graph it is not surprising that we also see those effects in the top layer as has also been shown in the original paper (Abnar and Zuidema, 2020).

3 Downstream Tasks

We also perform downstream task experiments on natural language inference and coreference resolution for both gendered and gender-neutral pronouns to investigate to what extent gender-neutral pronouns influence the performance.

Natural Language Inference Natural Language Inference (NLI) is commonly framed as a classification task, which tests a model's ability to understand entailment and contradiction (Bowman et al., 2015). Despite high accuracies achieved by SOTA models, we are yet to know whether they succeed in combating gender bias, especially in cross-lingual settings. We apply two multilingual models mBERT[5] (Devlin et al., 2019) and XLM-R[6] (Conneau et al., 2020) with cross-lingual fine-tuning, i.e., we fine-tune on English and apply both models also on Danish and Swedish. Therefore, mBERT was fine-tuned on the English MNLI train split and evaluated on XNLI. For XLM-R, we apply a model that has been fine-tuned on both MNLI and ANLI (Nie et al., 2020)[7]. For English we test both models on the MNLI test split, for Danish and Swedish we test on the extended XNLI corpus (Singh et al., 2019), the manual translation of the first 15000 sentences of the MNLI corpus (Williams et al., 2018) from English into 15 languages.

Coreference Resolution We also run pronoun resolution experiments on the Winogender dataset (Rudinger et al., 2018) where all 720 English sentences include an *occupation*, a *participant* and a *pronoun*. For each occupation, two similar sentences are composed, one where the pronoun refers to the occupation and one where it refers to the participant. Those sentences are then presented in versions with different pronouns (female, male, singular they). For our experiments, we compare performance for those pronouns and add a version for the gender-neutral pronoun *xe*. We run experiments with NeuralCoref 4.0 in SpaCy.[8]. For Danish, we apply the recently published coreference model (Barrett et al., 2021) to both the corresponding test set from the *Dacoref* dataset and a *gender-neutralized* version where we exchange gendered pronouns *hun/han* for either *høn* or singular *de*.[9]

4 Results

Natural Language Inference Accuracies for all languages and both models are displayed in Table 2. We overall see a very small drop in performance for the datasets with gender neutral pronouns compared to the original sentences. For mBERT we see differences of $0.09 - 1.43\%$, for XLM-R the drop

[5] `multi_cased_L-12_H-768_A-12`

[6] `xlm-roberta-large`

[7] https://huggingface.co/vicgalle/xlm-roberta-large-xnli-anli

[8] https://github.com/huggingface/neuralcoref

[9] So far, no Swedish coreference model has been published, we therefore leave this analysis for future work.

	en			da			sv		
	orig.	they	xe	orig.	de	høn	orig.	de	hen
mBERT	**83.33**	83.23	81.82	**71.15**	71.24	69.72	**71.91**	71.14	71.06
XLM-R	**95.13**	94.81	94.05	**80.19**	79.18	75.48	**78.79**	78.5	78.58

Table 2: Accuracy [in %] on NLI for English, Danish and Swedish for both models mBERT and XLM-R. Accuracies are calculated on the subset of sentences that contain relevant pronouns (924 for en and 2339 for da/sv). The first column for each language shows the accuracy on the original data, second and third columns show accuracies for respective gender-neutral pronouns. Please note, the total number of label flips in both directions for different pronouns is higher than the performance difference for all pair-wise comparisons. A baseline analysis where we exchanged punctuation ("." for "!") yields similar deviations from the original dataset than the changing pronouns.

is slightly higher with $0.21 - 4.71\%$. We see the biggest difference for the Danish pronoun *høn* in comparison to the original dataset.

	she	he	they	xe
acc in %	42.92	**43.75**	27.92	0

Table 3: Results for the pronuns resolution task on the English Winogender dataset.

	orig.	de	høn
F1-score	**0.64**	0.63	0.62
Prec.	**0.70**	0.69	0.69
Recall	**0.59**	0.57	0.56

Table 4: Results for the Danish coreference resolution task. Pronouns in the original dataset (orig.) have been exchanged for singular *de* and gender-neutral *høn*.

Coreference Resolution Table 3 shows accuracies on the English Winogender corpus for all four pronouns. We see a clear drop in performance from gendered pronouns (*she, he*) to both gender-neutral pronouns (*they, xe*). For *xe*, the model was not able to perform coreference resolution at all. In most cases it was not even recognized as part of a cluster and in the rare cases where it was, it was clustered with the wrong tokens. Please note that since this dataset is not labelled we are only classifying if the pronoun has been clustered with the correct entity. Results on the Danish Coref corpus, where we are able to perform a more extensive coreference resolution task are displayed in Table 4. We were able to replicate results from (Barrett et al., 2021) (the first column *orig.*). And see small drops in performance for singular *de* and *høn*.

5 Discussion

With this paper we provide a first study on how well language can handle gender-neutral pronouns in Danish, English and Swedish for various tasks. We observe an increase in perplexity for gender-neutral pronouns and correlations between perplexity on sentence level and attention flow on the pronoun, in particular for English and Swedish that gets stronger across layers. This indicates that language models indeed struggle with the use of gender-neutral pronouns, even with singular *they*, which has been used for many years as gender-neutral (Saguy and Williams, 2022). The reason for this most likely lies in the sparse representation of gender-neutral pronouns in the training data and the fact that language models, once they are trained and published usually are not updated (Bender et al., 2021). At the same time, we observe that word representations of all Swedish 3rd person pronouns grow closer in middle and top layers (see Figure 1) which suggests that relevant information is also learned for gender-neutral *hen*.

For NLI, we only see a small drop in performance when exchanging gendered pronouns for gender-neutral pronouns which is in the same range as a baseline analysis where we exchange punctuation ("!" for "."), except for Danish *høn*. We argue that classification in NLI probably does not heavily rely on individual pronouns in most cases. In stark contrast to pronoun resolution where we see a very clear drop in performance for English when applying singular *they* in comparison to both female and male pronouns, again this is surprising since in theory language models should have seen training samples where singular *they* has been used. The small drop in performance for Danish coreference resolution might be because this dataset does not solely focus on pronoun resolution, further investigation is needed here. We strongly argue that more needs to be done to adapt language models to a more gender inclusive language, initiatives like the rewriting task as proposed by Sun et al. (2021) need to be implemented and extended.

References

Irum Abbasi. 2020. The influence of technology on english language and literature. *English Language Teaching*, 13:1.

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Anonymous. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Submitted to ACL2022*. Under review.

Maria Barrett, Hieu Lam, Martin Wu, Ophélie Lacroix, Barbara Plank, and Anders Søgaard. 2021. Resources and evaluations for Danish entity resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 63–69, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marie Gustafsson Sendén, Emma A Bäck, and Anna Lindqvist. 2015. Introducing a gender-neutral pronoun in a natural gender language: the influence of time on attitudes and behavior. *Frontiers in Psychology*, 6:893.

Marie Gustafsson Sendén, Emma Renström, and Anna Lindqvist. 2021. Pronouns beyond the binary: The change of attitudes and use over time. *Gender & Society*, 35(4):588–615.

Laura Hekanaho. 2020. *Generic and Nonbinary Pronouns : Usage, Acceptability and Attitudes*. Ph.D. thesis, Helsinki, Finland.

Lisa Irmen. 2007. What's in a (role) name? formal and conceptual aspects of comprehending personal nouns. *Journal of Psycholinguistic Research*, 36(6):431–456.

Tore Kristiansen, Nikolas Coupland, Barbara Soukup, Sylvia Moosmüller, Frans Gregersen, Peter Garrett, Charlotte Selleck, Pirkko Nuolijärvi, Johanna Vaattovaara, Jan-Ola Ingemar Östman, Leila Mattfolk, Philipp Stoeckle, Christoph Hare Svenstrup, Stephen Pax Leonard, Kristján Árnason, Tadhg Ó Hifearnáin, Noel Ó Murchadha, Loreta Vaicekauskienė, Stefan Grondelaers, Roeland van Hout, Helge Sandøy, Mats Thelander, Elen Robert, Jannis Androutsopoulos, Peter Auer, Helmut Spiekermann, Allan Bell, Dirk Speelman, and Jane Stuart-Smith. 2011. Language change and digital media: A review of conceptions and evidence.

Sabrina J Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989.

Ehm Hjorth Miltersen. 2020. Singular de and its referential use in talk-in-interaction. *Scandinavian Studies in Language*, 11(2):37–37.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Theresa Redl, Stefan L Frank, Peter De Swart, and Helen De Hoop. 2021. The male bias of a generically-intended masculine pronoun: Evidence from eye-tracking and sentence evaluation. *PloS one*, 16(4):e0249309.

Jihyeon Roh, Sang-Hoon Oh, and Soo-Young Lee. 2020. Unigram-normalized perplexity as a language model performance measure with different vocabulary sizes. *arXiv preprint arXiv:2011.13220*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in

coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Abigail C Saguy and Juliet A Williams. 2022. A little word that means a lot: A reassessment of singular they in a new era of gender politics. *Gender & Society*, 36(1):5–31.

Anthony J Sanford and Ruth Filik. 2007. "they" as a gender-unspecified singular pronoun: Eye tracking reveals a processing cost. *Quarterly Journal of Experimental Psychology*, 60(2):171–178.

Jasdeep Singh, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. BERT is Not an Interlingua and the Bias of Tokenization. *The Workshop on Deep Learning for Low-Resource NLP at EMNLP 2019*.

Lydia Gabriela Speyer and Erik Schleef. 2019. Processing 'gender-neutral'pronouns: A self-paced reading study of learners of english. *Applied Linguistics*, 40(5):793–815.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Hellen P Vergoossen, Philip Pärnamets, Emma A Renström, and Marie Gustafsson Sendén. 2020a. Are new gender-neutral pronouns difficult to process in reading? the case of hen in swedish. *Frontiers in psychology*, 11:2967.

Hellen Petronella Vergoossen, Emma Aurora Renström, Anna Lindqvist, and Marie Gustafsson Sendén. 2020b. Four dimensions of criticism against gender-fair language. *Sex Roles*, 83(5):328–337.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

A Related Work

More eye-tracking studies have been conducted investigating the influence in processing cost for both gender-neutral pronouns and the generic male pronoun. Irmen (2007) and Redl et al. (2021) find male biases when using generic male pronouns in Dutch and generic role nouns in German. The authors of Sanford and Filik (2007) found a clear processing cost when using singular they in English, however their stimuli did not include any investigation of how (anti-)stereotypes influence this processing cost and is thus only in parts comparable to other studies. English datasets have been proposed to investigate gender bias in pronoun resolution but have not reported on performance differences between gendered and gender-neutral pronouns (Rudinger et al., 2018; Zhao et al., 2018; Webster et al., 2018). Sun et al. (2021) propose a rewriting task where data is transferred from gendered to gender-neutral pronouns to train more inclusive language models. The extensive survey on gender bias in NLP recently published by (Stanczak and Augenstein, 2021) also discusses research beyond binary gender formulation and the lacks of it.