Computer, en*hen*ce: POS-tagging improvements for nonbinary pronoun use in Swedish

Henrik Björklund[*]
he/him — han/honom
Umeå University
Umeå, Sweden
henrikb@cs.umu.se

Hannah Devinney[†]
they/them — hen/hen
Umeå University
Umeå, Sweden
hannahd@cs.umu.se

Abstract

Part of Speech (POS) taggers for Swedish routinely fail for the third person gender-neutral pronoun *hen*, despite the fact that it has been a well-established part of the Swedish language since at least 2014. In addition to simply being a form of gender bias, this failure can have negative effects on other tasks relying on POS information. We demonstrate the usefulness of semi-synthetic augmented datasets in a case study, retraining a POS tagger to correctly recognize *hen* as a personal pronoun. We evaluate our retrained models for both tag accuracy and on a downstream task (dependency parsing) in a classicial NLP pipeline.

Our results show that adding such data works to correct for the disparity in performance. The accuracy rate for identifying *hen* as a pronoun can be brought up to acceptable levels with only minor adjustments to the tagger's vocabulary files. Performance parity to gendered pronouns can be reached after retraining with only a few hundred examples. This increase in POS tag accuracy also results in improvements for dependency parsing sentences containing *hen*.

1 Introduction

The gender-neutral third person singular pronoun *hen* (subject/object form: *hen*; possessive form: *hens*) was added to the Swedish Academy's Glossary in 2015 (SAOL, 2015), following at least occasional use since the mid-20th century (Milles, 2013). The use and acceptance of *hen* has since increased (Gustafsson Sendén et al., 2021), although it remains much less common in media than *hon* ('she') or *han* ('he') (Svensson, 2021, 2022). Berglund (2022) provides a detailed study of the use of *hen* in blog posts from the years 2001–2017.

Despite its established history, Swedish Natural Language Processing (NLP) tools struggle to handle *hen* correctly, especially when compared to other pronouns. This is problematic both from a practical perspective (*hen* is increasingly used as a generic, e.g. on official forms) and from a bias perspective, as *hen* and other neopronouns are more likely to be used by gender minorities.

Part of Speech (POS) tagging is the task of assigning the individual words in a text to classes such as *noun*, *verb*, *pronoun*, etc. It is thus a fundamental task, one which many NLP systems rely heavily upon, e.g., systems for parsing, classification, translation, etc. This means that incorrect tagging may lead to errors in later steps. As an example, if *hen* is tagged as a noun, a translation system may well translate it into a noun rather than a pronoun.[1]

Swedish is an medium-resourced language, both in terms of high-quality labeled linguistic data and available tools. The available annotated datasets are of limited size and for the most part somewhat aged. When it comes to modern data-intensive tools, there is a series of BERT models trained by the National Library of Sweden that are publicly available. In the near future, GPT-SW3, a series of GPT style LLMs is also expected to be publicly released. As a consequence, when processing Swedish, we have to rely on combinations of modern LLMs and more classical NLP pipelines.

Apart from the direct usefulness of a Swedish POS tagger that can correctly tag *hen*, we also believe that it can be of general interest to investigate how to retrain POS taggers for new words, without access to up-to-date annotated datasets, which are expensive and very rarely produced or updated.

[*]Supported by the Wallenberg AI, Autonomous Systems and Software Program through the NEST project *STING*.
[†]Supported by the Umeå Centre for Gender Studies

[1]"Hen" actually exists as a noun in Swedish; it is an archaic term for a whetstone, and extremely rare in modern Swedish.

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 54–61, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_008

1.1 Bias Statement

In the NLP literature, "bias" can refer to various concepts, and is often not well-defined (Blodgett et al., 2020). We consider the overarching concept of algorithmic "bias" as the concern for how power structures *manifest* in language technologies. Power structures are a way of theorizing the pattern of underlying or hidden power relations in society/ies. We draw from Patricia Hill Collins' *matrix of domination*, which "describes the overall social organization within which intersecting oppressions originate, develop, and are contained" (Collins, 2000, p. 228). This draws attention to the complex interactions of different pieces in the whole system, encompassing four domains of power specified by Collins: *structural* (organization: laws, policies, large-scale institutions), *disciplinary* (administration/implementation of those laws and policies), *hegemonic* (system and circulation of ideas, favoring dominant groups), and *interpersonal* (everyday life and individual experiences).

Language technologies can operate in and be affected by several of these domains. In the case of POS taggers, we can consider their regulation of which terms are tagged as pronouns to be part of the disciplinary domain; while the abstract concept of a "standard" language determining which words "count as" pronouns is part of the structural domain, reinforced by hegemonic beliefs about the value of standard language. When the output from POS taggers is passed into other parts of an NLP pipeline, such as dependency parsing, this disciplinary power and regulation of legitimacy is also passed on. When these tools are applied, they become part of the matrix of domination across multiple domains as part of their interactions with the world.

However, even if there are no significant or "material" downstream effects of these mistakes, they are in and of themselves harms. "Non-standardized" pronouns and neopronouns, which are often the pronouns chosen by nonbinary[2] people, are delegitimized by automatic tagging tools mislabeling them as anything-but pronouns. This contributes to erasure and feelings of invisibility, and perpetuates the idea that these pronouns are "fake" and people who use them are "incorrect" or do not belong.

[2]We use nonbinary as an umbrella term for anyone outside or between the "binary" genders of women and men.

2 Background

Since pronouns are a much smaller class than other parts of speech such as nouns or verbs, more-or-less perfect accuracy should be expected from taggers. Indeed, we find that for the Swedish gendered pronouns *hon* and *han*, 100% accuracy is achieved for both taggers investigated (§3.1).

Stockholm-Umeå Corpus. The Stockholm-Umeå Corpus[3] (SUC) is an annotated corpus of texts from the 1990s (Gustafson-Capková and Hartmann, 2006). It contains about a million annotated words and is freely available for research purposes from Språkbanken (after signing a license agreement). The latest version (V3) was released in 2012.

efselab. The `efselab`[4] (Efficient Sequence Labeling) package provides a sparse perceptron-based architechture for POS tagging and other NLP tasks. It aims at computational efficiency, while still delivering a high accuracy (Östling, 2018). Once trained, efselab tagging is deterministic. Apart from the software needed to train models, the GitHub distribution also contains a pre-trained pipeline for Swedish, including POS tagging, named entity recognition, and dependency parsing. This out of the box tagger was trained on SUC, and has thus never "seen" instances of *hen*.

spaCy. The `spaCy` package[5] has three pre-trained pipelines for Swedish, differing in their sizes: small (`sm`), medium (`md`), and large (`lg`). The models are trained on SUC, Universal Dependencies Swedish Talbanken and varying amounts of unlabeld text data collected between 2018 and 2021 (spaCy). It can thus be assumed to have had instances of *hen* in its unlabeled training data, but not in its labeled data.

KB-BERT. The `KB-BERT` POS tagger[6] is based on Kungliga Bibliotekets (The National Library of Sweden's) BERT model, fine-tuned using the SUC corpus. As with `spaCy`, it can thus also be assumed to have had instances of *hen* in its unlabeled, but not in its labeled, training data.

[3]`spraakbanken.gu.se/en/resources/suc3`
[4]`github.com/robertostling/efselab`
[5]`spaCy.io`
[6]`https://huggingface.co/KBLab/bert-base-swedish-cased-pos`

2.1 Related Work

Brandl et al. (2022) show that large language models perform worse for gender-neutral pronouns in Danish, English, and Swedish than for gendered pronouns, measured both with respect to intrinsic measures such as perplexity and on several downstream tasks.

There are a number of systems for dependency parsing for Swedish, and in principle any framework for dependency parsing can be trained on the existing Swedish treebanks, such as UD-Talbanken. The parser included in `efselab`'s Swedish pipeline, and that we use here, is a pre-trained version of MaltParser (Nivre et al., 2007).

Data augmentation strategies are well-established for mitigating (binary) gender-stereotypical associations in NLP tools such as coreference resolution (Lu et al., 2020; Zhao et al., 2018), natural language inference (Sharma et al., 2020), dialog generation (Dinan et al., 2020), and abusive language detection (Park et al., 2018). For a general overview of data augmentation in NLP, see Feng et al. (2021).

Rewriting texts for data augmentation is not always a straightforward task, as exchanging words may require updates to other parts of the sentence to maintain grammatical agreement. Sun et al. (2021) demonstrate an algorithm for replacing gendered personal pronouns with neutral singular *they* in English, and Zmigrod et al. (2019) and Jain et al. (2021) propose methods for data augmentation in languages with grammatical gender. As *hen* follows the same paradigm as its gendered counterparts (see section 3.2), we find that it is sufficient to use simple replace rules with limited manual inspection.

3 Method

3.1 Initial Evaluation

The pre-trained taggers were initially tested for overall accuracy on the SUC test set, and for pronoun-specific accuracy on the Swedish Winogender Dataset[7]. SweWinogender is a challenge set, developed for diagnosing gender bias in coreference resolution systems follows a Winograd-style schema (Hansson et al., 2021). It is useful because in our setting it has a balanced frequency of *hen*, *hon*, and *han*, and also a good mixture of objective, subjective and possessive forms. Thus we can directly compare the accuracy across the pronouns, while being able to rule out context as a cause of differences.

We test both for accuracy across all morphosyntactic feature tags and "POS accuracy" which is only concerned with the top-level POS tag. We only report POS accuracy for `KB-BERT`, as it does not provide other feature information. Table 1 shows the POS accuracy of each tagger on all forms of *hen*, *hon*, and *han* on SweWinogender.[8] Table 4 shows the overall accuracy and POS accuracy for each tagger.

`KB-BERT` shows the best performance for *hen* for SweWinoGender, identifying it as a pronoun in nearly all cases. It also has the best POS accuracy on the SUC test set. Thus, it initially seems like there is not much to improve: we do not make modifications to KB-BERT, but continue reporting its performance as a reference point throughout the paper.

Despite an initial ability to sometimes correctly tag *hen* in the Swedish Winogender set (Table 1), the overall accuracy of Swedish `spaCy` is substantially worse than `efselab` (Table 2). In fact, the `spaCy` accuracy is at a level that is nowadays unacceptable.

For these reasons, we focus on `efselab` in the rest of the paper, using it as a case study to investigate the effects of augmenting the training data of a relatively light-weight tagger with synthetic data in order to incorporate a new pronoun into its repertoire. We are interested both in how much synthetic data is needed in order for the model to perform as well for the the new pronoun as for the others and in whether the addition of synthetic data deteriorates the overall performance.

3.2 Augmented SUC

The SUC corpus does not contain any instances of *hen* as a pronoun. In order to have access to tagged sentences using *hen*, we extracted sentences from SUC that use binary personal pronouns and constructed copies, replacing the pronouns with *hen*. We only swap tokens when the associated gold-standard tag is PN (personal pronoun) or PS (possessive personal pronoun). This check is necessary

[7] `spraakbanken.gu.se/resurser/swewinogender` (SweWinogender v1.0)

[8] The `KB` tokenizer sometimes splits composite words into separate tokens. In these cases, we only consider the `KB-BERT` POS tagging of the stem, in order to have an equal number of tokens for each model. This holds for all tests presented in this article.

SweWinogender	*hen*	*hon*	*han*
efselab	0.0	1.0	1.0
spaCy-sm	0.0	1.0	1.0
spaCy-md	0.82	1.0	1.0
spaCy-lg	0.75	1.0	1.0
KB-BERT	0.99	1.0	1.0

Table 1: Pronoun POS accuracy for the different baseline POS taggers on the SweWinogender dataset, reported across all morphological forms of each third person personal pronoun (208 tokens considered for each pronoun).

SUC-test	**Accuracy**	**POS acc.**
efselab	**0.9696**	0.9780
spaCy-sm	0.8857	0.9159
spaCy-md	0.9179	0.9420
spaCy-lg	0.9243	0.9459
KB-BERT	N/A	**0.9930**

Table 2: Baseline accuracy scores for the SUC test dataset, containing 23319 tokens. Under "Accuracy" we report the accuracy for tagging with POS *and* morphological information. This does not apply to `KB-BERT`, as it does not produce morphological tags. Under "POS acc.", we report the accuracy of the POS tagging, disregarding morphological tags.

because *Hans* can be both a possessive pronoun ('His') and a proper name. The appropriate morphological form[9] of the pronoun is used, as shown in table 3. We also update the morphological tag, to indicate that *hen* may be either the subject or object form. Capitalization is always preserved.

	hon	*han*	→	*hen*
subject	hon	han	→	hen
object	henne	honom	→	hen
possessive	hennes	hans	→	hens

Table 3: Replacement rules for singular personal pronouns in our en*hen*ced SUC.

Sentences where the replacement resulted in either *hen eller hen* ('ze or ze') or *hen och hen* ('ey and ey') required manual checking and correction. There were less than 50 of these instances in total. In all cases of *hen eller hen*, the original sentence was expressing a generic she-or-he, meaning the whole phrase could be collapsed into *hen*. For some cases of *hen och hen* no correction was required, e.g. in cases where the conjunction connects separate clauses. For the remaining sentences, a binary-gendered pronoun was re-introduced for clarity.

This resulted in 11 370 sentences using *hen*. We performed an 80/10/10 train/dev/test split on these sentences. This left us with a training set of 9 096 available sentences. For training, we combined this with the SUC training set in different proportions. Using 227 *hen* sentences makes the ratio of *hen* about 2% of the gendered pronouns in the resulting training set. This number was picked as a reasonable estimate of actual usage in modern Swedish (see, e.g., (Svensson, 2021, 2022)). To investigate whether less common pronouns need to be "over-represented" (compared to an approximated "realistic" usage) in training data to be correctly tagged, we also used training sets augmented with 10% (1 137) and 80% (9 096) of our total *hen* sentences, taken only from the training set.

3.3 Retraining

An `efselab` tagger contains two parts: the actual tagger and a statistical model trained on the training data. When the tagger part is built, it is provided with data files to build a vocabulary, with corresponding POS tags and morphological information. In order for the tagger to recognize *hen* as a pronoun, it is not sufficient to just train the statistical model on data containing examples of *hen*. The files that are used to build the vocabulary must be modified.

We thus trained five `efselab` models. The **baseline** model (`baseline`) is trained on SUC, using unmodified vocabulary files. The **mod. vocab** model (`hen0`) is trained on SUC, using modified vocabulary files. The three **en*hen*ced** models (`hen2, hen10, hen80`) are trained on SUC augmented with the given percentage of *hen* sentences, using modified vocabulary files.

4 Evaluation and Results

4.1 Part of Speech Tagging

We evaluated the models for accuracy based both on the full tags which include morphological information ("Accuracy") as well as the bare part of speech tags ("POS acc."). Because *hen* can be used as both subject and object form, our replacement strategy required some adjustment before both of these scores were brought into alignment. Two test datasets of comparable size, unseen in the training of any of the models, are used. The SUC test

[9] Although the object form of *hen* may also be written *henom*, we did not include this as it is not in common usage.

dataset is provided in SUC version 3.0, and is used unchanged. The *hen* test dataset is produced from the SUC test and development sets following the modification strategy described above. The results from these datasets are reported in table 4 and 5, respectively.

We evaluated the `efselab` models by providing the tokenized test sets as input and directly comparing the output to the SUC gold standard.

4.2 Dependency Parsing

We use the Swedish annotation pipeline provided in `efselab` to perform dependency parsing, with the default parsing model. This pipeline makes use of MaltParser (Nivre et al., 2007) (version 1.9.0), which incorporates POS information as a feature. Thus, we expect improvement in token-level accuracy for *hen* tokens.

Because SUC is not annotated with dependency information, we use the Swedish UD-Talbanken treebank[10] and evaluate on both the provided test set (UD test) and a smaller 'UD-HEN' test set consisting of only sentences that have been augmented in the same way as SUC. We report Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS), and Label Accuracy (L_{ACC}) on a token level, in Tables 6 and 7.

5 Discussion

Our initial findings showed that two common POS taggers for Swedish either cannot identify *hen* as a pronoun at all, or identify it at notably lower rates than other pronouns. This likely has downstream consequences on performance of language technologies relying on these taggers, and on the level of the taggers themselves is a problem for gender equality. It also demonstrates a weakness of such taggers, namely their ability to be flexible in light of language shift.

In our initial tests with SweWinogender, `KB-BERT` performed nearly-perfectly for *hen*: it only missed the two instances where *hens* was the first word of a sentence (and thus capitalized). However, SweWinogender is a very regularized test set (as it is designed for challenging coreference systems, not POS taggers), and `KB-BERT`'s performance for *hen* drops to less than 95% when tested on the more complex, realistic '*hen*' SUC test dataset. This makes it plausible that having only unlabeled data might not be sufficient to learn, e.g., pronouns that have recently come into use and are underrepresented in the data. As the `KB-BERT` model was originally fine-tuned for POS tagging on SUC, it seems reasonable that fine-tuning on the en*hen*ced SUC data could mitigate this weakness. Another reason for keeping tools such as `efselab` around is that at present is that the `KB-BERT` model does not provide more complex morphological information, which is desirable in some cases.

Training existing architectures on augmented data containing even a small number of sentences containing the pronoun *hen* can effectively correct for this disparity. We reach complete parity to binary-gendered pronouns, at 100% accuracy, with a representative sample (`hen2`), and see no real improvement when adding more sentences containing *hen* (`hen10` and `hen80`). This suggests that an up to date annotated dataset, based on contemporary Swedish usage, would be enough to obtain inclusive results, without the need for synthetic data, at least for the case of *hen*.

In terms of effect on downstream tasks, this improvement carries over to label accuracy for dependency parsing on *hen* tokens, with no loss of to LAS over all tokens. As nouns and pronouns often occupy similar grammatical roles, it is somewhat unsurprising that there is not also an effect on head accuracy (as measured by UAS).

Limitations

The current study only addresses one, relatively established, new personal pronoun in Swedish, and only pursues serious improvements to one tagger. Due to the under-resourced status of Swedish NLP, we only demonstrate the effects of this improvement on one "out of the box" downstream task. In future work, we hope to test these effects on other tasks prone to gendered biases, such as coreference resolution.

Although we find our strategy of re-training on augmented data to show good results for `efselab`, which is relatively lightweight, in general this type of constant re-training is not energy efficient, and therefore not environmentally responsible. Language, particularly inclusive language, is constantly shifting, meaning that more work of this type is inevitable to keep up with linguistic change. In future work, rule-based or other lightweight al-

[10] https://github.com/UniversalDependencies/UD_Swedish-Talbanken

SUC-test	Accuracy	POS Accuracy
baseline	0.9703	0.9786
hen0	0.9703	0.9786
hen2	0.9683	0.9771
hen10	0.9687	0.9774
hen80	0.9686	0.9774
KB-BERT	N/A	0.9929

Table 4: Results for the SUC test dataset, containing 23319 tokens, across different `efselab` models. `KB-BERT` is provided as a reference value for POS accuracy.

HEN-test	Accuracy	POS acc.	Hen acc.	Hen POS acc.
baseline	0.9093	0.9116	0.0000	0.0000
hen0	0.9775	0.9798	0.8870	0.8870
hen2	0.9941	0.9948	1.0000	1.0000
hen10	0.9945	0.9952	1.0000	1.0000
hen80	0.9953	0.9958	1.0000	1.0000
KB-BERT	N/A	0.9886	N/A	0.9408

Table 5: Results for the '*hen*' SUC test dataset, containg 20437 tokens (of which 1554 are *hen* or *hens*), across different `efselab` models. `KB-BERT` is provided as a reference value for POS accuracy.

UD-test	LAS	UAS	L_{ACC}
baseline	0.6087	0.6608	0.7565
hen0	0.6087	0.6609	0.7565
hen2	0.6108	0.6640	0.7571
hen10	0.6127	0.6655	0.7560
hen80	0.6068	0.6600	0.7544

Table 6: Word-level scores on the UD test set, containing 20386 tokens, across different `efselab` models.

UD-HEN-test	LAS	UAS	L_{ACC}	Hen LAS	Hen UAS	Hen L_{ACC}
baseline	0.6373	0.6860	0.7802	0.6534	0.7045	0.8068
hen0	0.6438	0.6891	0.7895	0.6932	0.7216	0.8807
hen2	0.6451	0.6906	0.7864	0.6932	0.7216	0.8920
hen10	0.6559	0.0707	0.7957	0.6989	0.7273	0.9034
hen80	0.6485	0.6947	0.7880	0.7045	0.7216	0.9091

Table 7: Word-level scores on the '*hen*' UD test set, containing 22778 tokens (of which 1266 are *hen* or *hens*), across different `efselab` models.

ternatives for updating models would be more desirable as solutions, or else combining many changes into one update to minimize retraining.

Further, our augmentation strategy is not well-suited for languages with different grammatical features from Swedish. Although Swedish does have grammatical noun classes, the (socially) gendered pronouns *han* and *hon* do not require agreement with any other terms in a sentence, meaning that we can replace them quite freely with relatively simple rules. This would not be the case in grammatically-gendered languages, such as French, Russian, or Hindi. To follow French as an example, if we would like to replace the binary pronouns *il* ('he') and *elle* ('she') with a neopronoun such as *iel*, we would first need to know whether a given instance refers to a person (replace) or an object (do not replace), and then would also need to update other terms in the text such as adjectives and pronouns to maintain grammatical agreement with the new pronoun. Alternative approaches, such as those described by Zmigrod et al. (2019) and Jain et al.

(2021) are required for the data augmentation.

Ethics Statement

Following the recommendations in Blodgett et al. (2020), we provide a full bias statement in section 1.1 detailing the risks we are trying to mitigate. Although gender is a sensitive attribute, we work at a level of abstraction (identifying POS information) that means our data does not contain *personal* identifying or sensitive information.

Due to the licensing requirements of SUC, we cannot distribute our training or test data. However, we release our modification code[11], meaning that anyone with access to SUC can themselves recreate the data, and even modify it for new pronouns.

Acknowledgements

The authors would like to warmly thank Robert Östling for prompt and helpful answers regarding the use of `efselab` and Jenny Björklund for helpful discussions and proof reading. The first author was partially funded by the Wallenberg WASP NEST STING project. The second author was co-funded by the Umeå Centre for Gender Studies.

References

Märta Berglund. 2022. *Hens* väg in i svenskan: En diakron korpusstudie av bruket av *hen* i bloggtexter. Master's thesis, Uppsala Unversity.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 3624–3630.

Patricia Hill Collins. 2000. *Black Feminist Thought*. Routledge, New York, New York, USA.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Sofia Gustafson-Capková and Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*.

Marie Gustafsson Sendén, Emma Renström, and Anna Lindqvist. 2021. Pronouns Beyond the Binary: The Change of Attitudes and Use Over Time. *Gender and Society*, 35(4):588–615.

Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender Dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 452–459, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating Gender Augmented Data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online. Association for Computational Linguistics (ACL).

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender Bias in Neural Natural Language Processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th*, 1 edition, pages 189–202.

Karin Milles. 2013. En öppning i en sluten ordklass? den nya användningen av pronomet hen. *Språk & Stil*, 23:107–140.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Robert Östling. 2018. Part of speech tagging: Shallow or deep learning? *Northern European Journal of Language Technology*, 5(1):1–15.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2799–2804.

SAOL. 2015. Svenska akademins ordlista 14.

[11] Link redacted for anonymous review.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2020. Evaluating Gender Bias in Natural Language Inference. In *NeurIPS 2020 Workshop on Dataset Curation and Security*.

spaCy. Available trained pipelines for swedish. https://spacy.io/models/sv. Accessed 2022-10-24.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English.

Anders Svensson. 2021. Hen ännu vanligare i svenska medier. https://spraktidningen.se/2021/01/hen-annu-vanligare-i-svenska-medier/. Accessed 2022-10-21.

Anders Svensson. 2022. Hen står still i svenska medier. https://spraktidningen.se/2022/01/hen-star-still-i-svenska-medier/. Accessed 2022-10-21.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1651–1661.