Improving Swedish Part-of-Speech Tagging for *hen*

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

Despite the fact that the gender-neutral pronoun *hen* was officially added to the Swedish language in 2014, state of the art part of speech taggers still routinely fail to identify it as a pronoun. We retrain both `efselab` and `spaCy` models with augmented (semi-synthetic) data, where instances of gendered pronouns are replaced by *hen* to correct for the lack of representation in the original training data. Our results show that adding such data works to correct for the disparity in performance.

1 Introduction

Part of Speech (POS) tagging is the task of automatically marking a term with its associated part of speech (i.e. whether it is a noun, verb, adjective, etc.). POS taggers are typically one of the earliest steps applied when solving natural langauge processing (NLP) tasks, and accurate tagging is essential for many 'downstream' tasks such as coreference resolution and parsing. Large language models have been shown to 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 (Brandl et al., 2022).

The taggers of two common pipelines (`efselab` and `spaCy`) are investigated. This work is partially motivated by experience in previous research (Devinney et al., 2020a,b), where we had to work around the fact that the tagger consistently mislabeled *hen*.

We find that current POS taggers struggle to handle non-standard pronouns. Since such pronouns are much more likely to be used by gender minorities, we can consider this disproportionate failure rate a form of gender bias. To correct for this, we retrain POS taggers on an augmented version of the Stockholm-Umeå corpus, where sentences containing binary personal pronouns *hon* or *han* have been copied and the pronouns replaced by *hen*. We will make available the most successful taggers.

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 "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 Partricia 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.

"Non-standard" pronouns, which are often the pronouns chosen by nonbinary[1] people, are delegitimized by automatic tagging tools mislabeling them as anything-but pronouns. This contributes to erasure and feelings of invisibility, and perpetuates

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

the idea that these pronouns are "fake" and people who use them are "incorrect" or do not belong.

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.

Stockholm-Umeå Corpus. The Stockholm-Umeå Corpus[2] (SUC) is a million-word collection of annotated Swedish texts from the 1990s (Gustafson-Capková and Hartmann, 2006). Version 3.0, released in 2012, improves the existing annotations and adds 7 million words of unannotated texts. It is freely available from Språkbanken for research purposes, after signing a licence agreement.

efselab. The `efselab`[3] (Efficient Sequence Labeling) package provides a sparse perceptron-based architechture for POS tagging and other NLP tasks, aimed at being computationally efficient, while still delivering a high accuracy (Östling, 2018). The distribution also provides a pre-trained pipeline for Swedish NLP tasks, including POS tagging. Users with a sufficiently large corpus can also use `efselab` to train a new tagger.

spaCy. The `spaCy` package[4] has three pre-trained pipelines for Swedish, in different sizes (`sm` small, `md` medium, and `lg` large). The taggers are trained on data from the Stockholm-Umeå corpus (version 3.0); the Universal Dependencies Swedish Talbanken; and, for the medium and large pipelines, a mix of other unlabeled texts taken from the internet, collected between 2018 and 2021 (spaCy).

Hen. The gender-neutral third person singular pronoun *hen* was added to the Swedish Academy's Dictionary in 2015 (SAOL, 2015), following at least occasional use since the mid-20th century (Milles, 2013). The use and acceptance of *hen* has increased (Gustafsson Sendén et al., 2021), although it remains much less common in media than *hon* or *han* (Svensson, 2021, 2022).

3 Method

3.1 Initial Analysis

We performed some initial tests on the pre-trained taggers using the the Swedish Winogender Dataset[5], a diagnostic "challenge" set for identifying gender bias in coreference resolution systems which follows a Winograd-style schema (Hansson et al., 2021). These sentences are useful because they are formulaic and contain an even distributions of the pronouns *hen*, *hon*, and *han* as well as a decent mixture of subject, object, and possessive forms. This lets us directly compare accuracy rates between the three pronouns, while knowing that the context the pronouns appear in is not varying.

For each of the four POS taggers (`efselab` baseline, `spaCy` baseline-sm, `spaCy` baseline-md, and `spaCy` baseline-lg), we get the tagged versions of all 624 test sentences from the SweWinogender dataset. Then, we extract only the target terms, i.e. {*hen, hens, hon, henne, hennes, han, honom, hans*}, and their associated tags. Then for each pronoun category {*hen, hon, han*} we calculate the accuracy as the rate at which these terms are identified as pronouns.

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

Table 1: Pronoun POS accuracy for the different POS taggers on the SweWinogender dataset.

Table 1 shows the results of the initial accuracy tests. The POS tagging is, as should be expected, 100% accurate for the pronouns *han* and *hon*. The larger POS taggers, which were trained after 2012 and thus are likely to have "seen" *hen* during training, can identify *hen* as a pronoun some of the time, but far from always.

To attempt to improve this, we retrained both systems on augmented versions of SUC with varying amounts of synthetic examples containing *hen*.

Augmented SUC. The SUC corpus contains no uses of *hen* as a pronoun. To get access to tagged data using *hen*, we extracted sentences from SUC that use *hon* and *han*, and constructed copies, but with *hen* as the pronoun. This resulted in a training set with 9096 sentences using *hen*. 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.

[2] `spraakbanken.gu.se/en/resources/suc3`
[3] `github.com/robertostling/efselab`
[4] `spaCy.io`
[5] `spraakbanken.gu.se/resurser/swewinogender` (SweWinogender v1.0)

This number was picked as a reasonable estimate of actual usage in modern Swedish. To investigate whether less common pronouns need to be "over represented" in training data to be correctly tagged, we also used training sets with 10% (1137) and 80% (9096) *hen* sentences.

efselab. 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 `efselab baseline`, trained on SUC with unmodified vocabulary, and the `efselab hen 0, 227, 1137`, and `9096` models, trained on SUC augmented with the given number of synthezised sentences, with modified vocabulary.

spaCy. A `spaCy` tagger is trained within a pipeline that may contain other parts. It is possible to train a pipeline from scratch or by continuing training of an existing model on new, compatible data. Although `spaCy` does allow for pretraining on unannotated data, we did not do so in order to remain consistent with the `efselab` taggers. We use the three models available for Swedish as our baselines, (`spaCy baseline-sm`, `baseline-md`, and `baseline-lg`) and train four further models (`spaCy hen 0, 227, 1137`, and `9096`) from scratch on SUC augmented with the given number of synthesized sentences.

3.2 Evaluation

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."). Two test datasets of comparable size, unseen in the training of any of the models, are used. The SUC test 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 as above.

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

Unlike `efselab`, `spaCy` expects raw text input for processing. We transformed both test sets into such files by re-joining the tokens into sentences before feeding them into `spaCy`. Due differences in the tokenizer, there can be alignment errors between the output and the SUC gold standard. To account for this, we aligned every sentence and then rejected any tokens which did not have an exact match.

SUC-test	Accuracy	POS acc.
efselab baseline	0.9696	0.9780
efselab hen 0	0.9696	0.9780
efselab hen 227	0.9686	0.9776
efselab hen 1137	0.9691	0.9775
efselab hen 9096	**0.9699**	**0.9784**
spaCy baseline-sm	0.8857	0.9159
spaCy baseline-md	0.9179	0.9420
spaCy baseline-lg	**0.9243**	**0.9459**
spaCy hen 0	0.9097	0.9488
spaCy hen 227	0.9183	0.9555
spaCy hen 1137	0.9144	0.9535
spaCy hen 9096	0.9180	0.9565

Table 2: Accuracies for the SUC test dataset, containing 23319 tokens.

4 Results

Table 2 shows the accuracy over all tokens for the SUC test dataset. Because the SUC test data does not contain any instance of the word *hen*, we do not report *hen* accuracy in this table. We note that the the `efselab hen 9096` tagger, which has been traied on a little more data does marginally better than the other `efselab` taggers. We also note that none of the `spaCy` taggers reach even 95% accuracy for POS tagging. The results for the best overall `efselab` and `spaCy` models are bolded.

Table 3 shows the accuracy over our *hen* test set, where every sentence contains an instance of *hen*, both over all tokens and for *hen* tokens. Since *hen* is such a common word in this set, the taggers that do not recognize it do extremely poorly. Notably, other taggers do better on this set than on the SUC test set, presumably because of typically simpler sentence structure, and the high frequency of personal pronouns. Again, the `efselab hen 9096` tagger has the best overall performance, but for the pure POS tagging task, the difference is marginal. The other `efselab` models struggle with the morphological information for *hen*, however. We conjecture that this is due to having seen

HEN-test	Accuracy	POS acc.	Hen acc.	Hen POS acc.
efselab baseline	0.9125	0.9135	0.0	0.0
efselab hen 0	0.9845	0.9890	0.9196	0.9964
efselab hen 227	0.9857	**0.9911**	0.9471	0.9994
efselab hen 1137	0.9857	0.9902	0.9639	0.9994
efselab hen 9096	**0.9867**	0.9899	**0.9813**	**1.0**
spaCy baseline-sm	0.7903	0.8204	0.0	0.0167
spaCy baseline-md	0.8794	0.9067	0.5100	0.5982
spaCy baseline-lg	0.8886	0.9108	0.5570	0.6233
spaCy hen 0	0.8998	0.9062	0.0	0.0039
spaCy hen 227	0.9814	0.9940	0.8738	0.9994
spaCy hen 1137	**0.9803**	**0.9938**	**0.8835**	**1.0**
spaCy hen 9096	0.9819	0.9930	0.8995	0.9987

Table 3: Accuracies for the 'hen' test dataset, containg 20437 tokens.

too few examples in training; see Section 5.1.

5 Discussion

Our initial findings showed that 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 consquences 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.

Training existing architectures on augmented data containing even a small number of sentences containing the pronoun *hen* can effectively correct for this disparity. This suggests that there is no need to over-represent gender-diverse language in datasets to obtain inclusive outcomes, at least for the task of part of speech tagging.

5.1 Limitations

The model trained with only 2% of the gendered pronouns being forms of *hen* struggles with the morphological information for *hen*, raising the question whether a frequency of *hen* proportional to common usage is insufficient. We conjecture that this is not the case, rather that due to the limited size of the SUC corpus, the tagger has not seen enough examples. Lacking access to a larger annotated corpus, we cannot test this conjecture.

Our method for producing synthetic data needs to be refined. The current version can, e.g., produce constructions such as "hen eller hen".

The `spaCy` framework is not designed for training from scratch on small corpora, so fine-tuning an existing model may very well yield better results.

5.2 Future Work

We will make our updated `efselab` model publicly available once our augmentation strategy has been refined. If a fine-tuned version of `spaCy` yields better results, we would also like to release that as a resource.

The current study only addresses one relatively established new personal pronoun in Swedish. Potential future work includes a comprehensive, multilingual analysis of how POS taggers treat neopronouns; as well as experimenting to see if this semi-synthetic data augmentation strategy could be effective in other contexts. As this type of constant re-training is not energy efficient, and therefore not environmentally responsible, rule-based or other alternatives for updating models in light of language shift would be more desirable as solutions.

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.

References

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.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020a. Crime and relationship: Exploring gender bias in NLP corpora. In *The Eigth Swedish Language Technology Conference*, SLTC.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020b. Semi-supervised topic modeling for gender bias discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (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.

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

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

SAOL. 2015. Svenska akademins ordlista 14.

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

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.