Is there Gender Bias in Dependency Parsing? Revisiting “Women’s Syntactic Resilience”

Paul Stanley Go[1] and **Agnieszka Falenska**[1,2]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
[1]paulstanleygo@yahoo.fr [2]agnieszka.falenska@ims.uni-stuttgart.de

Abstract

In this paper, we revisit the seminal work of Garimella et al. (2019), who reported that dependency parsers learn demographically-related signals from their training data and perform differently on sentences authored by people of different genders. We re-run all the parsing experiments from Garimella et al. (2019) and find that their results are not reproducible. Additionally, the original patterns suggesting the presence of gender biases fail to generalize to other treebanks and parsing architectures. Instead, our data analysis uncovers methodological shortcomings in the initial study that artificially introduced differences into female and male datasets during preprocessing. These disparities potentially compromised the validity of the original conclusions.

1 Introduction

NLP tools are commonly trained on textual corpora with authorship imbalances. For instance, since journalists are predominantly male[1], corpora derived from newspaper articles are largely written by men (Falenska et al., 2018; Garimella et al., 2019). Similarly, Wikipedia, a major resource for training NLP models (Devlin et al., 2019; Webster et al., 2019), is edited by a predominantly white and male group of contributors (Lam et al., 2011; Collier and Bear, 2012). This lack of diversity among authors can diminish the representation of minority voices (Bender et al., 2021) and lead to models that inherently mirror demographic imbalances (Hovy et al., 2020).

Garimella et al. (2019) was among the first to demonstrate how authorship imbalances can affect foundational NLP tasks like part-of-speech tagging and dependency parsing.[2] The authors trained models on sentences authored by females and males, observing error disparities in their results.[3] They found that models trained on male-authored sentences performed best on male test data, whereas models trained on a gender-balanced dataset yielded better results on female test data. These findings led them to conclude that sentences written by women exhibit greater “diversity” and complexity, which are better captured when training data includes contributions from both genders. In contrast, sentences by men showed less syntactic variability, resulting in decreased performance when female-authored sentences were included in the training set. Due to the heavy gender imbalance in the dataset (1:3 female to male authors), the authors concluded that the syntax of sentences written by women showed resilience despite the allocation bias, while men “lucked out” by having more training examples to boost accuracy.

The findings of Garimella et al. (2019) brought attention to the problem of gender bias in NLP models. The work was widely cited, for example, in the following work by the same authors (Garimella et al., 2021), influential surveys (Stanczak and Augenstein, 2021; Blodgett et al., 2020; Shah et al., 2020), and most importantly, as an argument that gender bias exists in NLP on the grammatical level (Lauscher et al., 2022). However, despite its significance, the study has notable deficiencies. Its scope is *limited to English* and only on *a single parsing architecture*. Moreover, the evaluation methodology *lacks any report of statistical significance testing* on the results. Given the minor differences in the obtained accuracy and the non-deterministic nature of neural models (Reimers and Gurevych, 2018), there is a potential that the findings of Garimella et al. (2019) could be attributed to chance.

To advance our understanding of potential gen-

[1] https://www.statista.com/statistics/625775/gender-news-reporting-us/

[2] Dependency parsing is the task of identifying the grammatical relationships between words in a sentence to form a syntactic dependency tree.

[3] We refer to Shah et al. (2020) for an overview of different types of biases.

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 269–279
August 16, 2024 ©2024 Association for Computational Linguistics

der biases in foundational NLP tasks such as part-of-speech tagging and dependency parsing, it is crucial to establish a well-defined foundation. Therefore, in this paper, we revisit Garimella et al. (2019) and aim to answer three research questions:

RQ1 Are the results presented in Garimella et al. (2019) reproducible and statistically significant?

RQ2 Do Garimella et al.'s (2019) results generalize to other languages and parsing architectures?

RQ3 What other factors, if not gender bias, could have been captured by their work?

We begin by replicating Garimella et al.'s (2019) methodology and rerunning their experiments (§3). Interestingly, our findings do not support the original claims regarding biases (§4). Further tests on the generalizability of these claims to a different language and parsing architecture also fail to replicate the original patterns. Our data analysis uncovers a small yet significant methodological flaw in the original study that can be responsible for the original results (§5). Consequently, we urge the gender bias research community to approach the results of Garimella et al. (2019) with caution. Moving forward, we recommend focusing more on specific syntactic differences related to demographic variations and their impact on model performance rather than relying solely on average scores, which can be misleading.

2 Bias Statement

According to the predictive bias framework proposed by Shah et al. (2020), the gender bias discussed in this paper is a form of selection bias – effects from the compositions of training data and their influence on downstream tasks. This selection bias manifests as error disparity, where models perform inconsistently across data from different demographic groups. While our focus is on dependency parsing, it is challenging to identify immediate, concrete harms directly caused by this bias. However, any subsequent applications that rely on these dependency parsers, such as authorship profiling based on syntactic trees (Morales Sánchez et al., 2022), could be affected. Depending on the specific application of the downstream task, this could lead to allocation or representation harms, where one demographic group might be unfairly treated or misrepresented due to biased model performance (Blodgett et al., 2020).

For our experiments, we require sentences annotated with the gender of their authors, along with gold-standard syntactic trees. To the best of our knowledge, we use the only two treebanks available that meet these criteria. These datasets categorize gender in binary terms, limiting our analysis to female and male authors. We recognize that this limitation excludes non-binary individuals, contributing to recognition bias against them.

3 Experimental Setup

We extend the experimental framework from Garimella et al. (2019) by incorporating additional data, parsing architectures, and robust evaluation.

3.1 Data

We use two well-established treebanks in English and German.

English We use the same gender-annotated subset of Penn Treebank (Marcus et al., 1993) as Garimella et al. (2019). It contains 19,399 trees for sentences from male authors and 7,282 for female.

German To compare the English results with a different language, we use the TIGER 2.2 treebank (Brants et al., 2004) comprised of syntactically-annotated German sentences from newspapers. A subset of the data was further annotated with the author's name and binary female/male gender by Falenska et al. (2018). The gender information was induced from the gold-standard morphological features of the authors' names. After removing all of the sentences annotated with HEADER and META labels, indicating meta-level information such as the article's title or time of document's creation, we were left with 3,550 trees for sentences written by female authors and 15,184 by male.

3.2 Preprocessing

Both English and German datasets are imbalanced wrt. to the gender of the authors. We will refer to these original datasets as RAW and use their BALANCED versions for the parsing experiments. For the balancing, we follow the same exact steps as Garimella et al. (2019):

1. Sort the sentences of each gender class in descending order according to the number of tokens.

2. Match each female sentence with a male sentence where the amount of tokens does not differ by more than 15%.

3. If there are no more male sentences that satisfy condition 2, the next male sentence in descending order with 5 to 30 tokens is chosen.

Once we have female and male datasets with an equal amount of sentences, we randomly choose an equal amount of sentences from those two to create a mixed-gender dataset of the same size. While Garimella et al. (2019) use 5-fold cross validation on their data for training and testing, we instead opt for the standard practice of a simpler 80-10-10 ratio split into training, development, and test sets when training models.

3.3 Dependency Parsers

Dependency parsers can generally be categorized into two classes: graph-based (Eisner, 1996; McDonald et al., 2005) and transition-based (Yamada and Matsumoto, 2003; Nivre, 2003). Since parsers from the two paradigms make different types of errors (McDonald and Nivre, 2007), we use one model from each category to additionally control for the role of the parsing architecture in our results.

Transition-based (TB) The original results of Garimella et al. (2019) used a transition-based parser SyntaxNet (Andor et al., 2016). However, the tool has been deprecated since the release of TensorFlow 2.0 in 2019.[4] Therefore, we re-implement all of their architecture with PyTorch (Paszke et al., 2019).[5]. Concretely, we use the arc-standard decoding algorithm (Nivre, 2004), Chen and Manning's (2014) feature function with fastText word vectors (Grave et al., 2018), and a feed-forward neural network with a ReLU activation function. We provide all the additional details and hyperparameters in Appendix A.1.1.

Graph-based (GB) In order to present a fair comparison to our transition-based parser, we use a graph-based parser with a similar neural architecture. We re-implement Pei et al.'s (2015) neural graph-based parser with Eisner's (1996) decoder, an adaptation of the Chen and Manning's (2014) architecture to a graph-based system. For more details and hyperparameters, we refer to Appendix A.1.2.

3.4 Evaluation

We evaluate the experiments using Unlabeled (UAS) and Labeled Attachment Score (LAS).[6] We use the three training sets to train FEMALE, MALE, and GENERIC models (to differentiate data from models, we will refer to the latter with capitalized names) and select the best-performing models based on the LAS of the corresponding development set. Subsequently, we test each of the models on the female, male, and generic test sets. We evaluate the statistical significance of all our models by following the recommendations of Reimers and Gurevych (2018): we train six models with different random seeds for each dataset and perform a Wilcoxon signed-rank test.

4 Parsing Results

We start by answering **RQ1** – are the results from Garimella et al. (2019) reproducible? For easier comparison, we repeat the original findings in Table 1a. The highest scoring models are highlighted in bold. The table presents the main finding of the study, namely that the GENERIC model performs the best on the female data and the MALE model on the male sentences.

4.1 English Results

We apply our TB parser to the English data, replicating the conditions used in Garimella et al. (2019). Table 1b presents the results averaged across six runs. The highest scores (i.e., the best LAS and UAS in the row) are highlighted in bold. Additionally, we report statistical significance for these results using superscripts with names of the models compared to which significance was achieved. For example, a $\mathbf{86.84}^{\mathrm{M}}$ UAS for the FEMALE model on the female test set not only indicates the highest score on this dataset compared to the MALE score (86.17) and GENERIC (86.69) but also signifies that the result is statistically significant relative to the MALE model, though not to the GENERIC.

Comparing Tables 1a and 1b, we observe that the patterns are markedly different. In our analysis, the FEMALE model achieves the best results on the female data, and the GENERIC model excels on both the male and generic data. Moreover, the statistical significance of the results is mixed, with some instances showing significance but not consistently across all results or in comparison to both other models. The only consistent finding with Garimella et al.'s (2019) is that the MALE model performs better on male sentences than the FEMALE model, as indicated by the $^{\mathrm{F,M}}$ significance.

[4] It would not be possible to run SyntaxNet without installing TensorFlow 1.x and all its associated old dependencies, making it impractical to run on modern systems.

[5] The code is available at https://github.com/paulstanleygo/goparser

[6] The percentage of tokens that received the correct head and label (LAS) or just head (UAS).

Train / Test	Female LAS	Male LAS	Generic LAS
female	83.17	83.12	**83.46**
male	81.15	**83.21**	82.53
generic	82.01	**83.11**	83.03

(a) Results reported by Garimella et al. (2019)

Train	Female		Male		Generic	
Test	UAS	LAS	UAS	LAS	UAS	LAS
female	**86.84^{M}**	**85.24**	86.17^{F}	84.58	86.69	85.11
male	$84.73^{M,G}$	83.03^{G}	85.39^{F}	83.70	**85.46^{F}**	**83.73^{F}**
generic	85.39	83.76	85.39	83.83	**85.74**	**84.09**

(b) Averages across six runs with different random seeds. Statistical significance is shown with a superscript indicating the models with which the significance is associated.

Table 1: Transition-based (TB) test results for English. Highest performing models are highlighted in bold (separately for UAS and LAS).

However, this is only observed in the UAS metric.

Interestingly, one additional pattern emerges from the analysis – sentences written by female authors are the easiest to parse. Regardless of the model used, all achieve the highest UAS and LAS scores on this dataset. Conversely, sentences authored by males prove to be the most challenging, consistently showing the lowest scores. We will explore this finding in the later discussion.

4.2 German Results

We switch to **RQ2** and ask whether the results from Garimella et al. (2019) can be replicated in a different language and parser architectures. Table 2 presents the German test results from TB and GB, averaged across six models. For TB (Table 2a), unlike the English results, we observe some similarities to the findings of Garimella et al. (2019). The Generic model achieves the highest scores on the female dataset, and the Male model surpasses the others on the male (UAS) and generic datasets (both metrics). However, none of these differences are statistically significant, except for the performance of Generic compared to Female on the male test set – a result that is not relevant for the narrative of Garimella et al.'s (2019).

Switching to GB (Table 2b), we observe that the performance differences are not consistent across parsing architectures. Interestingly, the results show more parallels with the English TB, where the Female model performs best on the female data, and the Generic model excels on the male and generic data. The statistically significant results also align more closely with the TB English results. Most importantly, these findings are similarly inconsistent with Garimella et al. (2019).

Finally, across both parsing architectures, the same clear pattern emerges as for the English results: sentences written by female authors are the easiest to parse, while those authored by males are the most difficult.

4.3 Error Analysis

The results presented in Tables 1 and 2 do not confirm the findings from Garimella et al. (2019). However, since UAS and LAS average scores across all dependency arcs, there might be still some patterns that we do not observe by only looking at single numbers. Therefore, as a final sanity check, we zoom into these results by performing error analysis on the models' performance. Following McDonald and Nivre (2011, 2007), we look at dependency length and distance to root to determine if there are any differences in parsing errors between models trained on the different data.

Figure 1 presents a sample of the results – the TB performance on the female and male datasets.[7] We select these datasets because they are crucial for the scenarios highlighted by Garimella et al. (2019), i.e., Generic on female data and Male on the male data. We leave the other results to Appendix A.2 together with analysis of distance to root, which shows similar patterns to the dependency length. Overall, the results corroborate our averaged findings. For the female dataset (left), up to a dependency length of 9, the Female model performs the best, followed by the Generic and then the Male model. Beyond this length, the differences vary, likely due to the limited number of long arcs. For the male dataset (right), Male slightly outperforms the others for arcs up to a length of 3, but thereafter is outperformed by the Generic model. In conclusion, we do not find indicators that would align with the results of Garimella et al. (2019).

[7] We analyze models with the highest validation LAS.

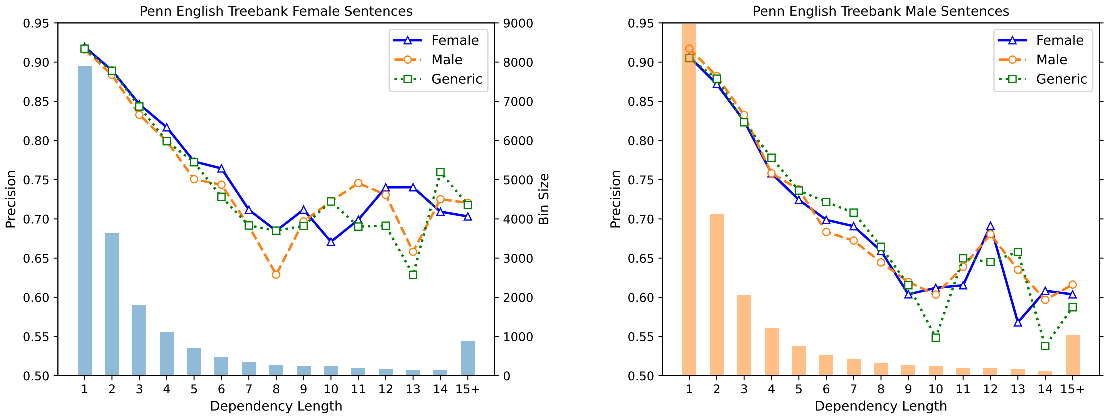
Train	Female		Male		Generic	
Test	UAS	LAS	UAS	LAS	UAS	LAS
female	80.52	76.67	80.70	76.66	**80.82**	**77.10**
male	77.25^{G}	73.02	**78.06**	73.89	78.00^{F}	**74.01**
generic	79.45	75.34	**80.19**	**76.14**	79.93	75.94

(a) Transition-based parser (TB)

Female		Male		Generic	
UAS	LAS	UAS	LAS	UAS	LAS
80.28^{M}	**75.72**	79.81	75.04	80.03	75.30
77.35^{M,G}	72.36^{M}	78.07^{F}	**73.37**^{F}	**78.13**^{F}	73.31
78.50	73.53	78.58	73.66	**78.74**	**73.75**

(b) Graph-based parser (GB)

Table 2: German test results averaged across six runs with different random seeds. Highest performing models are highlighted in bold (separately for UAS and LAS). Statistical significance is marked with a superscript indicating the models with which the significance is achieved.



(a) Female sentences

(b) Male sentences

Figure 1: TB precision for the English datasets relative to dependency length.

5 Data Analysis

If not gender bias, what was captured by the models of Garimella et al. (2019)? To answer **RQ3**, we perform analysis of our training datasets.

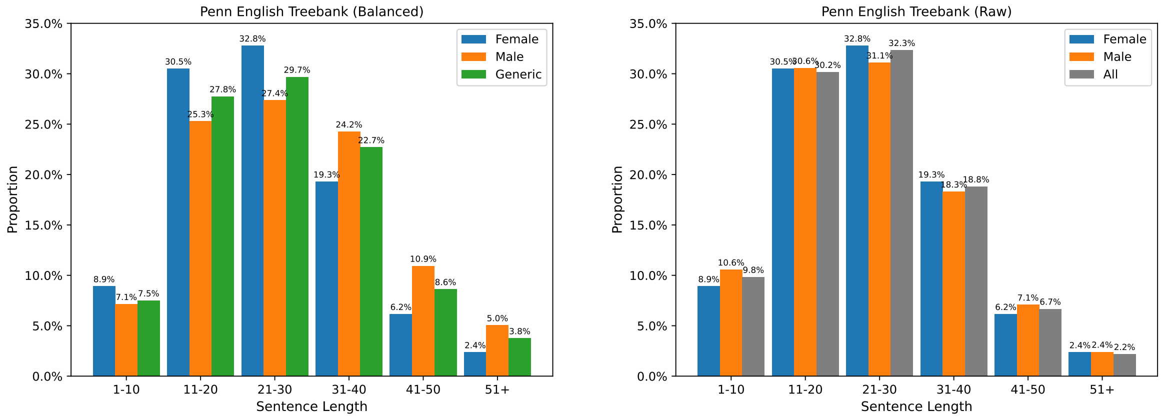
5.1 Sentence Length

We begin by examining the most straightforward factor – sentence length. Figure 2 displays the English data divided into bins by the number of tokens for the three datasets: female, male, and generic. The results for BALANCED (Figure 2a), the dataset that we used for training all the parsers, reveal a distinct pattern: female-authored sentences are shorter, with more falling within the 11-20 and 21-30 length bins. In contrast, male-authored sentences are more frequently in the longer 31-40, 41-50, and 51+ bins. Given that parsing accuracy generally declines with increased sentence length McDonald and Nivre (2011, 2007), this result can explain the pattern that we consistently observed across languages and architectures, i.e., that female sentences are "easier" to parse than male.

The results from Figure 2a exhibit the opposite trend from what is generally assumed in the previous literature, that female sentences are typically longer (Cornett, 2014, among others). However, as shown Figure 2b, this finding can not be attributed to the sociolinguistic factors in the data, but simply Garimella et al.'s (2019) preprocessing steps described in Section 3.2. In the original RAW dataset, male sentences are slightly more frequent in the 1-10 and 41-50 length categories, while female sentences predominate in the 21-30 and 31-40 ranges, with the 11-20 range being roughly equivalent for both genders. The balancing procedure used by Garimella et al. (2019) alters this distribution, resulting in shorter female sentences and longer male sentences. Originally, the average RAW male sentences were 0.24 tokens shorter than those of females in English and 0.13 tokens shorter in German. After preprocessing, the average length of BALANCED male sentences became 3.19 tokens longer than female sentences in English and 2.41 tokens longer in German. As a result and by accident, all the parsing results were influenced.

5.2 Tree Characteristics

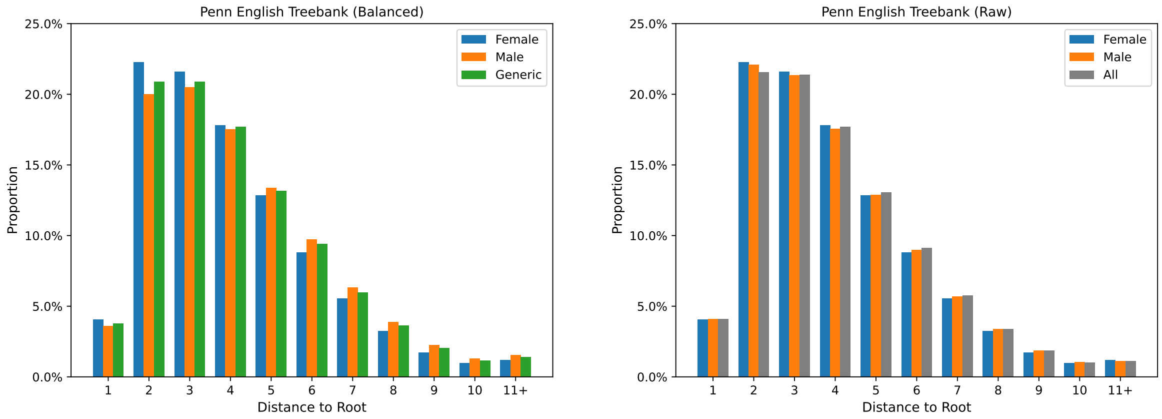
Dependency parsing is a structure prediction task where the number of tokens in sentences is strongly related to other treebank characteristics, such as



(a) BALANCED dataset

(b) RAW dataset

Figure 2: Proportion of sentences with different lengths in the English datasets.



(a) BALANCED dataset

(b) RAW dataset

Figure 3: Proportion of distance to root lengths in the English datasets.

the types and configurations of arcs in the trees. Therefore, by modifying the distribution of sentence lengths, it is possible to impact many other attributes of the tree structures. Figure 3 illustrates the proportions of distances to the root in both the RAW and BALANCED English datasets. In the RAW datasets (Figure 3b), there are no major differences in distance to root between genders. However, looking at the BALANCED datasets (Figure 3a), we see a different distribution. There are more tokens with distance to root of 1 to 4 in the BALANCED female dataset and conversely, more tokens with distance to root of 5 to 11+ in the BALANCED male dataset. This demonstrates that the balancing procedure results in a shorter average distance to root in the female dataset and a longer average distance to root in the male dataset.[8] Given that arcs further from the root are typically more challenging to parse (McDonald and Nivre, 2011), this provides another insight into why all models consistently show lower performance on the male-authored datasets.

6 Conclusion

In this paper, we revisited the seminal work on gender bias by Garimella et al. (2019). Our analysis demonstrated that their findings do not generalize to other languages or parsing architectures and, more critically, are not reproducible even with the same parsing architecture and dataset as the original study. A consistent observation from our work was that sentences written by females were easier to parse than those written by males. However, this pattern was due to a methodological oversight in the original study, where the preprocessing step inadvertently produced longer male sentences. As sentence length correlates with more complex tree structures, such as long arcs and dependents far

[8] A similar pattern for dependency length is less pronounced and visible only for arcs of 15+ tokens (see Figure 6 in Appendix A.2).

from the root, this error introduced artificial parsing difficulty. Coupled with our inconsistent statistical significance results across various applications, these findings challenge the validity of the gender bias claims made by Garimella et al. (2019).

7 Acknowledgements

This work was supported by the Ministry of Science, Research, and the Arts, Baden-Württemberg through the project IRIS3D (Reflecting Intelligent Systems for Diversity, Demography and Democracy, Az. 33-7533-9-19/54/5).

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. 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.

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.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation*, 2:597–620.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Benjamin Collier and Julia Bear. 2012. Conflict, confidence, or criticism: An empirical examination of the gender gap in Wikipedia contributions. In *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012*, pages 383–392. ACM.

Hannah E Cornett. 2014. Gender differences in syntactic development among English speaking adolescents. *Inquiries Journal/Student Pulse*, 6(03).

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.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Agnieszka Falenska, Kerstin Eckart, and Jonas Kuhn. 2018. Moving TIGER beyond sentence-level. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online. Association for Computational Linguistics.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You sound just like your father" Commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Shyong K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren G. Terveen, and John Riedl. 2011. WP:Clubhouse? An exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, 2011, Mountain View, CA, USA, October 3-5, 2011*, pages 1–10. ACM.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 91–98. Association for Computational Linguistics.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131, Prague, Czech Republic. Association for Computational Linguistics.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Damián Morales Sánchez, Antonio Moreno, and María Dolores Jiménez López. 2022. A white-box sociolinguistic model for gender detection. *Applied Sciences*, 12(5).

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve Restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA. Omnipress.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the Eighth International Workshop on Parsing Technologies (IWPT*, pages 149–160, Nancy, France.

Joakim Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pages 50–57, Barcelona, Spain. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2015. An effective neural network model for graph-based dependency parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 313–322, Beijing, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2018. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *CoRR*, abs/1803.09578.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 896–903, New York, NY, USA. Association for Computing Machinery.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, Italy. Association for Computational Linguistics.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333, Beijing, China. Association for Computational Linguistics.

Hiroyasu Yamada and Yuji Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 195–206, Nancy, France.

A Appendix

A.1 Parsing Hyperparameters

A.1.1 Transition-based parser

In general, we re-implement the SyntaxNet architecture. We incorporate Weiss et al.'s (2015) refinements to the Chen and Manning (2014) architecture by replacing the nonlinear activation function with ReLU (Nair and Hinton, 2010) and increasing the number of hidden layers to two. We apply dropout (Srivastava et al., 2014) to the hidden layers and following Kiperwasser and Goldberg (2016), we also add a word dropout that is inversely proportional to the frequency of the word to better deal with out-of-vocabulary words. The word embeddings are initialized with pre-trained 300-dimensional fastText word vectors (Grave et al., 2018), while all other weights are randomly initialized with a Kaiming uniform distribution (He et al., 2015). We purposely refrain from using more expressive feature representations such as the BiLSTM feature extractor (Kiperwasser and Goldberg, 2016) since there is a possibility that the increased expressiveness may influence our gender bias results and make it difficult to compare with Garimella et al.'s (2019) results. Moreover, for simplicity, we exclude SyntaxNet's beam search since it is used for alleviating search error and omitting it is unlikely to affect the overall result concerning gender bias. Table 3 summarizes all the details and used hyperparameters.

Decoder	Arc-standard
Word embedding dimension	300
Part-of-speech embedding dim.	32
Dependency label embedding dim.	32
Number of hidden layers	2
Hidden layer dimensions	256, 256
Hidden layer dropout p	0.5
Word dropout α	0.25
Word embedding initialization	fastText
Weight initialization	Kaiming uniform
Criterion	Cross-entropy loss
Optimizer	Adam
Learning rate	1e-5
nonlinear activation function	ReLU

Table 3: Hyperparameters for TB.

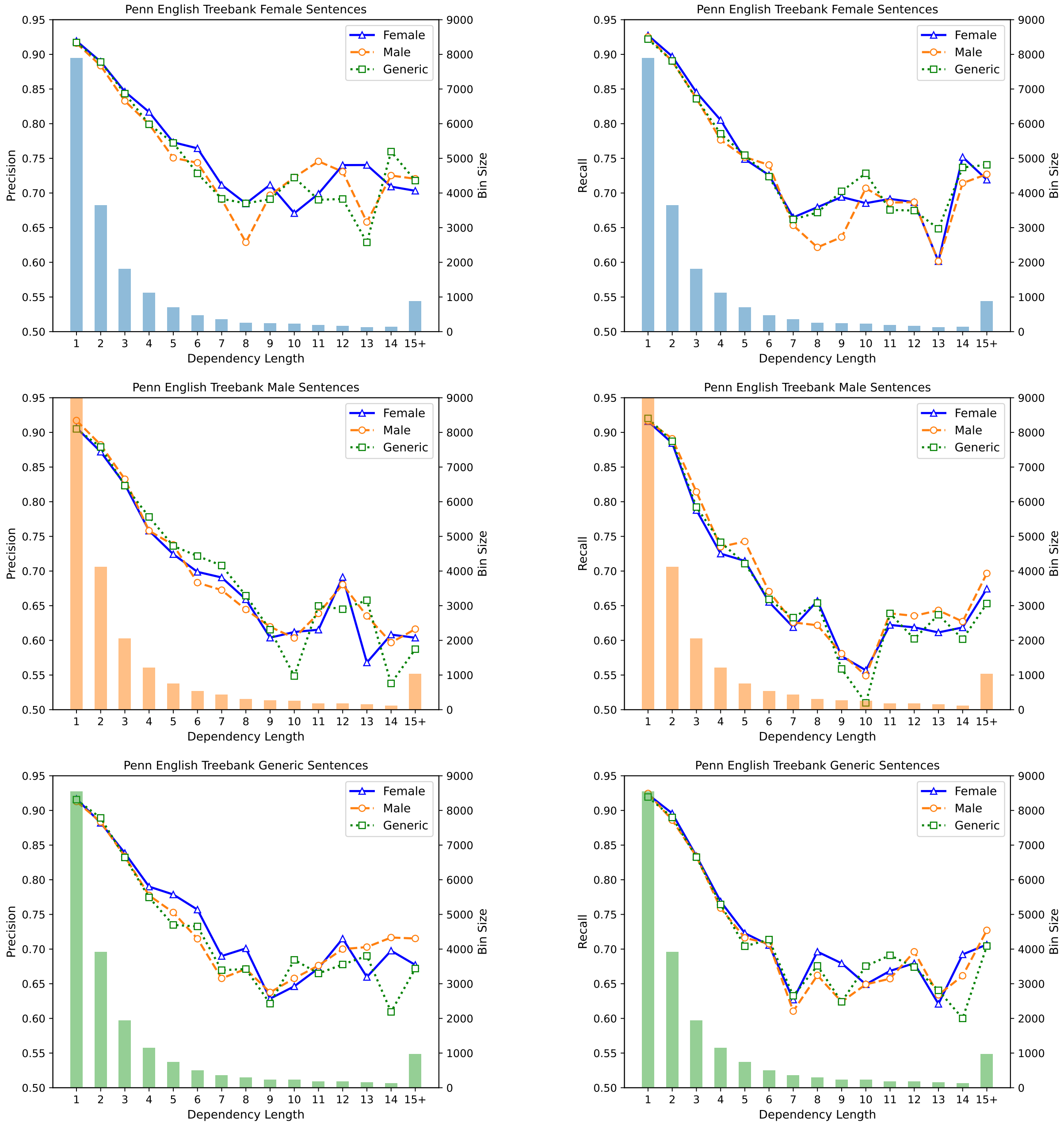
A.1.2 Graph-based parser

We use two hidden layers to match our transition-based parser and follow Kiperwasser and Goldberg (2016) in adding word dropout and using loss augmented inference (Taskar et al., 2005) by augmenting the scores of all incorrect arcs with a constant value of 1. The word embeddings are initialized with pre-trained 300-dimensional fastText word vectors (Grave et al., 2018), while all other weights are randomly initialized with a Xavier uniform distribution (Glorot and Bengio, 2010). Once again, we refrain from using more expressive feature representations for comparison purposes and use Pei et al.'s (2015) *1-order-atomic* features. Hyperparameters can be found in Table 4.

Decoder	Eisner's
Word embedding dimension	300
Part-of-speech embedding dimension	32
Distance embedding dimension	32
Number of hidden layers	2
Hidden layer dimensions	256, 256
Hidden layer dropout p	0.5
Word dropout α	0.25
Word embedding initialization	fastText
Weight initialization	Xavier uniform
Criterion	Hinge loss
Optimizer	Adam
Learning rate	1e-3
nonlinear activation function	Tanh-cube

Table 4: Hyperparameters for GB.

A.2 Parsing Results and Data Analysis



Penn English Treebank Female Sentences

Penn English Treebank Female Sentences

Penn English Treebank Male Sentences

Penn English Treebank Male Sentences

Penn English Treebank Generic Sentences

Penn English Treebank Generic Sentences

Figure 4: TB precision (left) and recall (right) on the English datasets relative to dependency length.

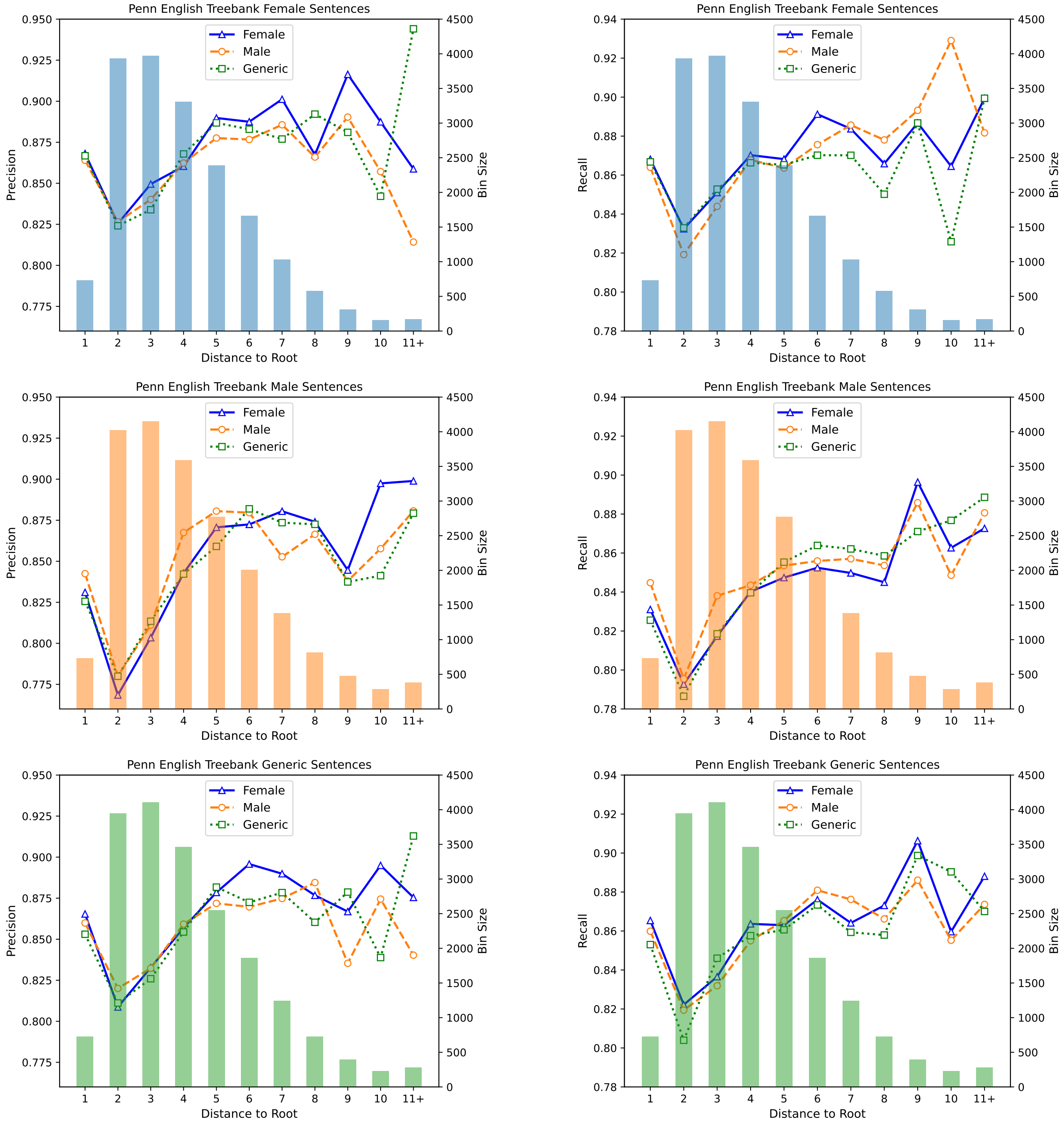


Figure 5: TB precision (left) and recall (right) on the English datasets relative to distance to root.

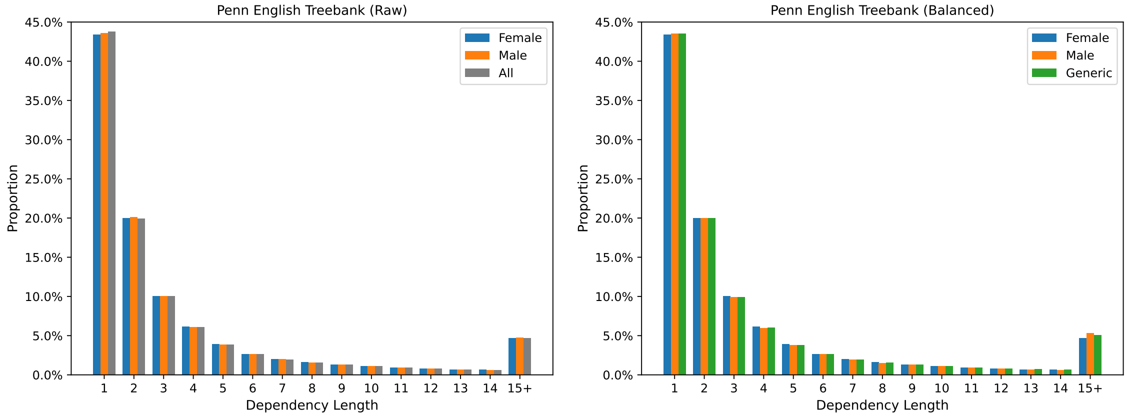


Figure 6: Proportion of dependency lengths in the RAW (left) and BALANCED (right) English datasets.