Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing

Aparna Garimella[1], **Carmen Banea**[1], **Dirk Hovy**[2], **Rada Mihalcea**[1]

[1]Computer Science and Engineering,
University of Michigan, Ann Arbor, MI
{gaparna,carmennb, mihalcea}@umich.edu

[2]Department of Marketing,
Bocconi University, Milan, Italy
dirk.hovy@unibocconi.it

Abstract

Several linguistic studies have shown the prevalence of various lexical and grammatical patterns in texts authored by a person of a particular gender, but models for part-of-speech tagging and dependency parsing have still not adapted to account for these differences. To address this, we annotate the Wall Street Journal part of the Penn Treebank with the gender information of the articles' authors, and build taggers and parsers trained on this data that show performance differences in text written by men and women. Further analyses reveal numerous part-of-speech tags and syntactic relations whose prediction performances benefit from the prevalence of a specific gender in the training data. The results underscore the importance of accounting for gendered differences in syntactic tasks, and outline future venues for developing more accurate taggers and parsers. We release our data to the research community.

1 Introduction

Sociolinguistic studies have shown that people use grammatical features to signal the speakers' membership in a demographic group, with a focus on gender (Vigliocco and Franck, 1999; Mondorf, 2002; Eckert and McConnell-Ginet, 2013). Mondorf (2002) shows systemic differences in the usage of various types of clauses and their positions for men and women, stating that women have a higher usage of adverbial (*accordingly*, *consequently*[1]), causal (*since*, *because*), conditional (*if*, *when*) and purpose (*so*, *in order that*) clauses, while men tend to use more concessive clauses (*but*, *although*, *whereas*). Similar results hold across various languages in Johannsen et al. (2015).

This correlation between grammatical features and gender has important ramifications for statistical models of syntax: if the training sample is unbalanced, these differences inadvertently introduce a strong gender bias into the training data. Such demographic imbalances are amplified by the model (Zhao et al., 2017), which in turn can be detrimental to members of the underrepresented demographic groups (Jørgensen et al., 2015; Hovy and Søgaard, 2015; Hovy and Spruit, 2016). Since several works use syntactic analysis to improve tasks ranging from data-driven dependency parsing (Gadde et al., 2010) to sentiment classification (Moilanen and Pulman, 2007; Socher et al., 2013), underlying model biases end up affecting the performance of a wide range of applications. While data bias can be overcome by accounting for demographics, and can even improve classification performance (Volkova et al., 2013; Hovy, 2015; Bolukbasi et al., 2016; Benton et al., 2017; Zhao et al., 2017; Lynn et al., 2017), there is still little understanding on the amount and sources of bias in most training sets.

In order to address gender bias in part-of-speech (POS) tagging and dependency parsing, we first require an adequate size data set labeled for a) *syntax* along with b) *gender* information of the authors. However, existing data sets fail to meet both criteria: data sets with gender information are either too small to train on, lack syntactic information, or are restricted to social media; sufficiently large syntactic data sets are not labeled with gender information and rely (at least in part) on news genre corpora such as the Wall Street Journal (WSJ). To address this problem, we augment the WSJ subset of the Penn Treebank corpus with gender, based on author first name. To our knowledge, this is the first work that explores syntactic tagging while accounting for gender.

[1] We exemplify in parentheses conjunctions or conjunctive adverbs that introduce and link in a subordinating relationship the given type of subordinate clause.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3493–3498
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

Contributions. The main contributions of this paper are as follows:

- We annotate a standard POS-tagging and dependency parsing data set with gender information.
- We conduct experiments and show the role played by gender information in POS-tagging and syntactic parsing.
- We analyze POS and syntactic differences related to author gender.

2 Annotating PTB for Gender

The Penn Treebank (Marcus et al., 1993) is the de facto data set used to train many of the POS taggers (Brill, 1994; Ratnaparkhi, 1996; Toutanova and Manning, 2000; Toutanova et al., 2003) and syntactic parsers (Klein and Manning, 2003; Nivre and Scholz, 2004; Chen and Manning, 2014). It contains articles published in the WSJ in 1989, as well as a small sample of ATIS-3 material, totalling over one million tokens, and manually annotated with POS tags and syntactic parse trees.

We supplement the WSJ articles with metadata from the ProQuest Historical Newspapers database, which indexes, among others, WSJ articles released between 1923 and 2000, and provides fields such as author names. Out of the original 2,499 WSJ articles, 1,814 are found in ProQuest and their metadata is retrieved. 556 articles with an empty $Author$ field are removed, resulting in 1,258 WSJ articles with author information. Using a combination of regular expressions and manual verification, we extract author names for 1,006 articles (the remaining 252 articles do not have actual author names).

We isolate the first names using regular expressions, and follow Prabhakaran and Rambow (2017) to automatically assign gender and compute a gender ambiguity score taking into consideration: (1) the list of first names obtained based on Facebook profiles by Tang et al. (2011); and (2) the Social Security Administration's (SSA) baby names data set.[2] The Facebook list has male and female assignment scores for each name, while the SSA maintains a data set of counts for baby names and gender for each year since the 1880s. If both databases agree in their gender assignment, we use that as the final label (987 articles). For the remaining 19, we manually identify the author gender by cross-referencing the names online. 5 of these only had a first name initial, and thus could not be resolved and were discarded. The gender mapping results in 1,001 gender tagged WSJ articles. Discarding 115 articles with joint authorship and considering only articles with both POS tags and parse trees results in a final set of **804 articles** from the Treebank.

The final set of articles includes 379 unique authors, with a heavy gender imbalance of 1 to 3 (96 female and 283 male). The total number of sentences in female articles is 7,282, with a mean of 21.17 tokens per sentence ($\sigma = 10.03$), while the male articles consist of 19,400 sentences, with a mean of 20.99 tokens per sentence ($\sigma = 10.52$). This is similar to the findings of Cornett (2014), who also notes a lengthier utterance mean for women versus men (her study focuses on adolescents).

We use the Universal Dependencies (UD) v1.4 (Nivre et al., 2016) annotation guidelines for parse trees and POS tags, and accordingly, convert the constituency trees from the Penn Treebank (PTB) format to the CoNNL format.[3] We then map the POS tags to the universal POS tag set.[4]

3 The Effect of Gender in POS Tagging and Dependency Parsing

To assess whether author gender affects parsing performance, we train the state-of-the-art transition-based neural network model SyntaxNet[5] (Andor et al., 2016) on the data (with default parameters), and test whether stratified training can alleviate these effects. We evaluate performance for individual POS-tags and dependency relations, as well as over all the tags and relations.

Stratifying the Training Data. Since the WSJ data has a heavy gender imbalance (1:3 female to male articles), we stratify the data by discarding male examples so that the number of female and male sentences and tokens do not differ by more than 15%: (1) We sort the female and male WSJ sentences in descending order of number of tokens. (2) For each female sentence F_i with f_i number of tokens, we select a male sentence M_j such that the number of tokens $m_j \in$

[2] http://www.ssa.gov/oact/babynames/limits.html

[3] https://nlp.stanford.edu/software/stanford-dependencies.shtml.

[4] The data sets are annotated with the 16 universal POS-tags; conj is used for both sconj and conj tags.

[5] https://github.com/tensorflow/models/tree/master/syntaxnet

$[0.75f_i, 1.25f_i]$. (3) If we run out of male sentences which qualify for this condition, we choose the next male sentence in descending order with number of tokens $m_j \in [5, 30]$. Table 1 shows the number of sentences and tokens in the WSJ data before and after balancing for gender.

We train the model in three scenarios: (1) on female data, (2) on male data, and (3) on generic data containing an equal number of male and female sentences. All three data sets have an equal number of sentences.

	RAW		BALANCED	
GENDER	SENT.	TOKENS	SENT.	TOKENS
FEMALE	7,282	175,107	7,282	175,107
MALE	19,400	461,742	7,282	202,144

Table 1: Number of sentences and tokens in the raw and balanced WSJ data.

Evaluation. We report standard evaluation metrics: accuracy (ACC) – the percentage of tokens that have a correct assignment to their part-of-speech (for part-of-speech tagging); and *labeled attachment score* (LAS) – the percentage of tokens that have a correct assignment to their heads *and* the correct dependency relation (Nivre et al., 2004) (for dependency parsing).

In each training setting, we generate five random training-test splits at a 90:10 ratio on the WSJ data set. In order to derive parameters for SyntaxNet, each train split is further randomly split into five folds. When creating the folds, we ensure that sentences authored by the same author are not shared across splits to avoid overfitting to the writing styles of individual authors, rather than learning the underlying gender-based differences as they pertain to syntax.

TRAIN:	GENERIC	FEMALE	MALE
TEST	POS ACCURACY		
GENERIC	95.81	95.49	95.74
FEMALE	**95.96**	95.90	95.47
MALE	95.47	95.03	**96.08**
	DEPENDENCY LAS		
GENERIC	83.03	82.01	83.11
FEMALE	**83.46**	83.17	83.12
MALE	82.53	81.15	**83.21**

Table 2: Results for part-of-speech tagging (ACC) and dependency parsing (LAS) on WSJ test data.

In each training scenario, we evaluate the models on: (1) female-only data, (2) male-only data, and (3) generic data containing an equal number of male and female sentences (364 sentences from each gender), such that all test settings share the same number of sentences (10% of 7,282 = 728). Since we have 5 test folds, and each fold in turn has 5 validation folds (for parameter tuning), we report results averaged over the 25 total runs to ensure robustness.

4 Results and Discussion

Table 2 (top) shows the POS-tagging accuracies for labeling the WSJ test data. We should note that while accuracy differences may be relatively small, they are within the margins of recent state-of-the-art improvements (Andor et al., 2016) in a task that achieves extremely high accuracy and where further improvement can only be incremental. Considering performance across the three different training scenarios, the female test data sees a slight benefit from a mixed training set, achieving its highest accuracy of 95.96%, while male test data only achieves the highest performance (96.08%) when training on male-only data, representing a relative error rate reduction of 13.46% when compared to the generic model.

The setting closest to current POS tagging setups is embodied by training on the generic model. In this case, the female test data achieves its highest accuracy (95.96%), but the male test data achieves only a second best performance (95.47%). This difference suggests an area of possible improvement in performance for off-the-shelf POS taggers.

We see a similar pattern in dependency parsing (Table 2, bottom), where the female test set achieves the highest LAS accuracy performance on the mixed training set (83.46%). The male test set obtains its highest accuracy when the training is performed on male-only data, with a relative error reduction of 3.89% as compared to training on generic data.

It seems that female writings are more diverse, with a complexity that can best be approximated with mixed-gender training samples. This setting improves performance by relative error reductions of (1.46%, 1.72%) (ACC, LAS) when compared to training on female-only data, and (10.82%, 2.01%) (ACC, LAS) when compared to training on male-only data. The male test sentences appear to display less variability, and therefore can-not benefit the same amount of information from the spectrum displayed by female training data; actually, any time female-authored sentences are present in the training set (whether as all female-data or generic data), performance drops for male test data.

When comparing male and female-only training sets and their ability to generalize to the opposite gender, we notice that male training data is more maleable and lends itself better to be used when testing on female samples, but not the reverse.

We note that the WSJ exemplifies a highly formal and scripted newswire genre, where gender differences are likely less pronounced, yet they still surface. We will likely observe even stronger language differences in a large, informal data set comprising both gender and syntactic information. These differences can be leveraged to achieve a better performance for core NLP tasks.

Train:	Generic	Female		Male	
	Acc	Acc	Err	Acc	Err
		Male Test			
noun	93.74	92.51	-19.63	**94.23**	7.92
det	99.09	99.09	-0.13	**99.13**	4.08
num	99.23	99.34	15.35	**99.35**	16.60
pron	99.17	99.11	-6.69	**99.19**	2.75
propn	93.97	90.10	-64.14	**95.26**	21.41
		Female Test			
pron	98.91	**99.12**	18.99	98.97	4.64
aux	98.60	**98.77**	12.12	98.39	-14.75
adj	92.12	**92.62**	6.37	92.36	3.06
propn	94.66	**94.97**	5.76	91.60	-57.33

Table 3: Tag-wise results for part-of-speech tagging on WSJ test data; Accuracies (Acc) and relative error reduction rates (Err) versus generic models are reported.

We also observe clear gender-based performance improvements at the tag level (Table 3). For instance, models trained on male-only data better predict `nouns`, `determiners`, `numerals`, `pronouns` and `proper nouns` for male test data, compared to models trained on mixed data (with a relative error rate reduction between 2.75% and 21.41%). Similarly, female-trained models better predict `pronouns`, `auxiliaries`, `adjectives`, and `proper nouns` for female test data, compared to models trained on mixed data (with a relative error rate reduction between 5.76% and 18.99%). For 8 out of the 16 POS tags, mixed training achieves best results for either female or male test data.

Train:	Gen.	Female		Male	
	Las	Las	Err	Las	Err
		Male Test			
csubj	25.20	27.89	3.60	**36.13**	14.61
iobj	47.11	40.61	-12.29	**48.59**	2.80
acl	63.93	60.47	-9.60	**66.09**	5.99
compound	75.06	72.95	-8.45	**77.26**	8.83
xcomp	74.39	72.26	-8.30	**75.38**	3.85
dobj	84.48	82.13	-15.17	**85.20**	4.66
conj	82.45	80.74	-9.77	**82.82**	2.11
nummod	92.00	91.24	-9.42	**93.08**	13.53
		Female Test			
amod	91.18	**91.46**	3.11	91.08	-1.18
cop	92.78	**93.89**	15.47	92.80	0.34
appos	79.44	**80.31**	4.21	80.13	3.38
cc:preconj	54.68	**65.09**	22.96	50.78	-8.60

Table 4: Tag-wise results for dependency parsing on WSJ test data; LAS and relative error reduction rates (Err) versus generic models are reported.

In dependency parsing (Table 4), models trained on female data better predict `amod`, `cop`, `appos`, and `cc:preconj` labels for female test sets (with a relative error rate reduction between 3.11% and 22.96% compared to generic models). Similarly, male-trained models are able to outperform mixed models on male test data for `csubj`, `iobj`, `acl`, `compound`, `xcomp`, `dobj`, `conj` and `nummod` with a relative error rate reduction between 2.11% and 14.61%. In dependency parsing, mixed training never achieves the best per tag results for either male or female test sets.

This suggests that leveraging the idiosyncrasies for specific tags displayed by each gender could help create gender-agnostic models that leverage the syntactic strengths of each gender, and improve prediction accuracy for both. It is to be noted that there is a heavy topic overlap between the male and female WSJ articles, with a Pearson correlation of 0.85 between the male and female topic distributions[6], indicating that the differences in performance between male and female models on various evaluation sets are not from topical shifts, but from syntactic variations.

5 Conclusion

Our experiments show that women's syntax displays resilience: POS taggers and dependency parsers trained on any data perform well when

[6] The topic distributions were extracted using Latent Dirichlet Allocation (Blei et al., 2003). We use the LDA implementation included with the Python Gensim library (Řehůřek and Sojka, 2010) with 10 topics.

tested on female writings. Male syntax, on the other hand, is parsed or tagged best when sufficient male-authored data is available in the training set. This suggests that men "lucked out" with respect to the gender imbalance in the WSJ training data: a more balanced or more female-heavy data set could have caused significant drops in the performance of automatic syntax analysis for male writers. The gender annotated WSJ data provides a starting point for leveraging gendered grammatical differences and the development of better and fairer models and tools for syntax annotation, as well as for the many NLP down-stream tasks that use syntax in their models.

The WSJ author gender information is publicly available from `http://lit.eecs.umich.edu/downloads.html`.

Acknowledgments

This material is based in part upon work supported by the Michigan Institute for Data Science, by the National Science Foundation (grant #1815291), and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Michigan Institute for Data Science, the National Science Foundation, or the John Templeton Foundation. The authors would like to thank the reviewers of the various drafts for their comments. Dirk Hovy is a member of the Bocconi Institute for Data Science and Analystics (BIDSA) and the Data and Marketing Insights (DMI) unit.

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)*, ACL 2016, pages 2442–2452, Berlin, Germany.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. In *Proceedings of the 15th European Chapter of the Association of Computational Linguistics (Volume 1: Long Papers)*, EACL 2017, pages 152–162, Valencia, Spain.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Quantifying and reducing stereotypes in word embeddings. *CoRR*.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI 1994, pages 722–727, Menlo Park, CA, USA.

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)*, EMNLP 2014, pages 740–750, Doha, Qatar.

Hannah E. Cornett. 2014. Gender differences in syntactic development among english speaking adolescents. *Inquiries Journal/Student Pulse*, 6(3):1–6.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Phani Gadde, Karan Jindal, Samar Husain, Dipti Misra Sharma, and Rajeev Sangal. 2010. Improving data driven dependency parsing using clausal information. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL HLT 2010, pages 657–660, Los Angeles, California, USA.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP 2015, pages 752–762, Beijing, China.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, ACL-IJCNLP 2015, pages 483–488, Beijing, China.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL 2016, pages 591–598, Berlin, Germany.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*, pages 103–112.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (Volume 1)*, ACL 2003, pages 423–430, Sapporo, Japan.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pages 1157–1166, Copenhagen, Denmark.

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

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, volume 7 of *RANLP 2007*, pages 378–382, Borovets, Bulgaria.

Britta Mondorf. 2002. Gender differences in English syntax. *Journal of English Linguistics*, 30(2):158–180.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning at HLT-NAACL 2004*, CoNLL 2004, Boston, Massachusetts, USA.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016, pages 1659–1666, Portoro, Slovenia.

Joakim Nivre and Mario Scholz. 2004. Deterministic dependency parsing of English text. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING 2004, Geneva, Switzerland.

Vinodkumar Prabhakaran and Owen Rambow. 2017. Dialog structure through the lens of gender, gender environment, and power. *arXiv preprint arXiv:1706.03441*.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 1 of *EMNLP 1996*, pages 133–142, Philadelphia, PA.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013, pages 1631–1642, Seattle, Washington, USA.

Cong Tang, Keith Ross, Nitesh Saxena, and Ruichuan Chen. 2011. Whats in a name: a study of names, gender inference, and gender behavior in facebook. *16th International Conference on Database Systems for Advanced Applications*, pages 344–356.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, NAACL-HLT 2003, pages 173–180, Edmonton, Canada.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, SIGDAT-EMNLP 2000, pages 63–70, Hong Kong, China.

Gabriella Vigliocco and Julie Franck. 1999. When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language*, 40(4):455–478.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, number October in EMNLP 2013, pages 1815–1827, Seattle, WA, USA.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pages 2979–2989, Copenhagen, Denmark.