Moving Away from Binary Gender in Corpora: Replacing “He/She” with the Singular Generic “They”

Jessica Cusi
Department of Linguistics
Georgetown University
jdc286@georgetown.edu

Abstract

This project has two components: a “Singular They” dataset (ST dataset) consisting of 429 sentences and a rule-based model (RB model) in which the dataset was tested on. The ST dataset is composed of instances where the binary gender pronoun form “he/she” are replaced with the singular generic “they” and corresponding verbs in four categories are adjusted for grammaticality. The four categories are: *be* verbs, *have* verbs, *do* verbs, and regular verbs ending with *-s*. When the RB model was evaluated, it received an accuracy score of 0.465. Limitations to the dataset are size and pronoun variation. Future directions include more data, an improved rule-based model, and potential Transformer models.

1 Introduction

The uses of the pronoun “they” have changed over the centuries. It is most commonly used as the third-person plural pronoun, such as in: *The children have recess at noon. They will eat lunch beforehand*. In this example, *the children* is the antecedent of *they*.

However, “they” is also used as a singular pronoun. Recently, the singular “they” has emerged as a personal pronoun for someone who identifies as non-binary, agender, or gender fluid. For example, in the case where *Alex* is non-binary and prefers they/them pronouns, the proper sentences would be: *Alex is studying computer science. They are learning how to program with Java*.

Nonetheless, this paper focuses on the second use of the singular “they,” which is “they” as the singular generic third-person pronoun. The singular generic “they” (SGT) is used to refer to an antecedent whose gender is unknown or irrelevant, or in cases where the gender must be concealed. Consider the [?] in the following: *The author wrote under a pseudonym. [?] published 3 novels in the last 5 years*. No information is given about the *author* in the first sentence. Therefore, one must decide which pronoun to use in place of the antecedent. The correct pronoun to replace [?] should be “they.” Thus, the sentences would be: *The author wrote under a pseudonym. They published 3 novels in the last 5 years*.

However, it is important to note that there have been many societal and linguistic arguments over the decision to use the SGT. In many genres of writing over the last few centuries, there have been different preferences for the pronoun that should be in place of [?], ranging from the generic masculine (GM) “he”, to binary variations such as “he/she,” “he or she,” or “(s)he,” and the singular generic “they.” There has been an abundance of research on the transition of English grammar—through a prescriptive perspective—to refer to an antecedent with an unknown or irrelevant gender (Mackay, 1980; Bradley, 2020). Historically, an ungendered antecedent was denoted with the GM, and studies like Foertsch and Gernsbacher (1997) have explored the cognitive implications of the masculine pronoun usage versus the singular generic “they.” Martyna (1980) argues that the generic “he” (GM) is a form of sexist language.

For more gender-inclusive language, SGT is the best option. Although the singular generic third-person pronoun “they” has been used in written text as early as the 14th century (Baron, 2018), it has not always been seen as popular and acceptable in formal and academic English. It was not the standard in academic writing for American Psychological Association (APA) style until 2019 (APA Style, 2022). Before 2019, the binary form “he/she” was preferred over “they.” That being said, there are still people who disagree with the use of SGT. Doll (2013) argues that its usage is indicative of sloppy writing and incorrect pluralization. Furthermore, some grammarians in the 18th century argued against SGT because it was incorrect that a plural pronoun (i.e., the original form of "they" as a third person plural pronoun) could take a singular antecedent (Baron, 2018).

Nevertheless, because English has historically used the singular generic "they" and has made it the preferred pronoun for referring to an unknown-gender antecedent in academic writing (APA Style, 2022), this should be the standard in datasets as well. Use of the singular generic they should be the best practice for artificial intelligence (AI) and machine learning (ML) models. However, since instances of binary gender pronouns are present in corpora rather than SGT, any model that is trained with this data will continue the cycle of using gender-exclusive language. This problem is what the current project attempts to address: the presence of "he/she" in current corpora, which goes hand-in-hand with the lack of the singular generic third person "they." Therefore, the "singular they" dataset, along with the rule-based baseline model created to make pronoun changes from binary pronouns to SGT, can be a beneficial tool for progress toward more gender-inclusivity AI and ML tools. The dataset and RB model will be discussed in more detail in later sections.

2 Related Work

2.1 Singular "They"

Many researchers have done more specific work on the singular generic third-person pronoun "they." Regarding the evolution of language, Balhorn (2004) maintains that internal developments in English, rather than social factors, have led to the usage of the singular generic "they."

In a different light, Miller & James (2019) investigated the effect of using the generic masculine pronoun on comprehension by replicating MacKay and Fulkerson (1979)'s experiments that tested whether or not students interpret the generic masculine "he" as truly generic. Their findings suggest that the pronominal dominance hypothesis (i.e., the lexical meaning of a pronoun determines the interpretation of its antecedent) and the pronominal surrogate hypothesis (i.e., the nature of an antecedent completely determines the interpretation of a pronoun)–both of which are defined by MacKay and Fulkerson (1979)–are complementary rather than competing and both influence language comprehension.

In an exploration of the different generic pronouns, LaScotte (2016)'s study on the use of SGT found that 79% of participants preferred to use "he or she" forms or the singular "they" over solely the generic masculine "he" or generic feminine "she" when referring to a singular genderless noun. Furthermore, within the "gender-inclusive approach" (i.e., "he or she" or singular "they" in this study), participants used SGT the majority of the time (68%).

Lastly, in her thesis, Anna Maryskova (2021) analyzes the singular generic "they" versus gendered pronouns in the Corpus of Contemporary American English (COCA) and British National Corpus to observe gender-neutrality in English. Though this current project uses data similar to Maryskova, it focuses on building a dataset and rule-based model to replace the binary gender "he/she" instances rather than analyzing the specific instances of SGT versus gendered pronouns.

2.2 Computational Approaches to Singular "They" and Gendered Pronouns

Based on the author's knowledge, there are no existing publications or computational models that have directly addressed or tackled the task of substituting the binary "he/she" for the singular generic "they" in corpora. A handful of previous research has been done regarding the gender bias in coreference resolution (Rudinger et al., 2018; Zhao et al., 2018). The most similar and recent research on the singular "they" was a coreference resolution task on "they," where Baumler and Rudinger (2022) found that current coreference models accept singular generic "they" more than "they" as a singular personal pronoun.

The new ST dataset and RB model presented in this paper attempts to fill this gap on exploring binary gender pronouns and SGT in ML applications. That being said, with the advancements of large language models (LLMs), some models already have the ability to generate sentences and stories using SGT based on the pure abundance of data they are trained on. Nonetheless, this new dataset and simple rule-based model that corrects binary pronouns to SGT can still be beneficial for fixing or replacing existing outdated datasets.

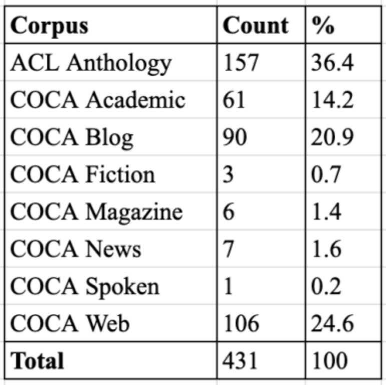
3 Data

Binary gender pronoun forms are not truly generic nor inclusive to people who do identify as non-binary, agender, or gender-fluid. Therefore, AI and machine learning models should not be trained with data that contains these forms (e.g., he/she, him/her, his/hers). To combat this issue, this “singular they” dataset (ST dataset) was created to include the singular generic “they” in replacement of the binary gender forms. As an attempt to align with dataset standards, the following sections answer some questions posed in “Datasheets for Datasets” by Gebru et al. (2021).

3.1 Data Collection & Composition

The data for the ST dataset was collected during March 2023 via Georgetown's CQP webpage.[1] Data was initially collected from 9 corpora: ACL Anthology and 8 genres of the Corpus of Contemporary American English (COCA) (Academic, Blog, Fiction, Magazine, News, Spoken, TV & Movies, and Web). In the end, however, the portion from the TV & Movies corpus was dropped because the data was too inconsistent and difficult to work with.

In the CQP interface, the following was used as the search query for every corpus: *[word = “he/she”]*. This query retrieved all instances of the binary form “he/she.” The other binary pronoun variations (e.g., him/her, his/hers) were excluded from the initial search. However, if any of these variations were included in the gathered data, they were replaced with their corresponding singular “they” forms (i.e., him/her → them/their, his/hers → theirs). After the query search, 20% of the sentences containing “he/she” in each corpus were randomly selected so that 573 total sentences would be collected.[2] In summary, using a stratified sampling method, the researcher compiled 429 sentences for the ST dataset. The number of sentences dropped from 573 to 429 after taking out the TV & Movies corpus data, as well as post-validation (discussed later) where some “final” sentences did not make sense to include anymore for various reasons. The breakdown of the dataset is shown in Figure 1.



Corpus	Count	%
ACL Anthology	157	36.4
COCA Academic	61	14.2
COCA Blog	90	20.9
COCA Fiction	3	0.7
COCA Magazine	6	1.4
COCA News	7	1.6
COCA Spoken	1	0.2
COCA Web	106	24.6
Total	431	100

Figure 1. “Singular They” dataset distribution.

For the 429 data points, each instance has an index, text ID, original sentence, and gold standard sentence. The index is the data point number; the text ID is the given filename of the document where the text is from (taken from CQP Web); the original sentence contains the raw sentence as extracted from CQP Web, and the gold standard sentence is the transformation of the original sentence. The transformation includes the replacement of the binary gender word “he/she” to “they,” along with any other corresponding binary gender forms that appear in the sentence. In the gold standard sentence, abnormal spacing, punctuation, and spelling (i.e., non-American English spellings) were corrected. An example of the dataset is shown in Figure 2.

Although the ST dataset was created solely by the author, the 429 sentences were validated by one peer. The validator was asked to ensure that the gold standard sentences (i.e., the sentences in the “gold-text” column) were grammatically correct, had the proper SGT pronouns and usages, and had no other minor errors (e.g., spacing and spelling).

3.2 Preprocessing

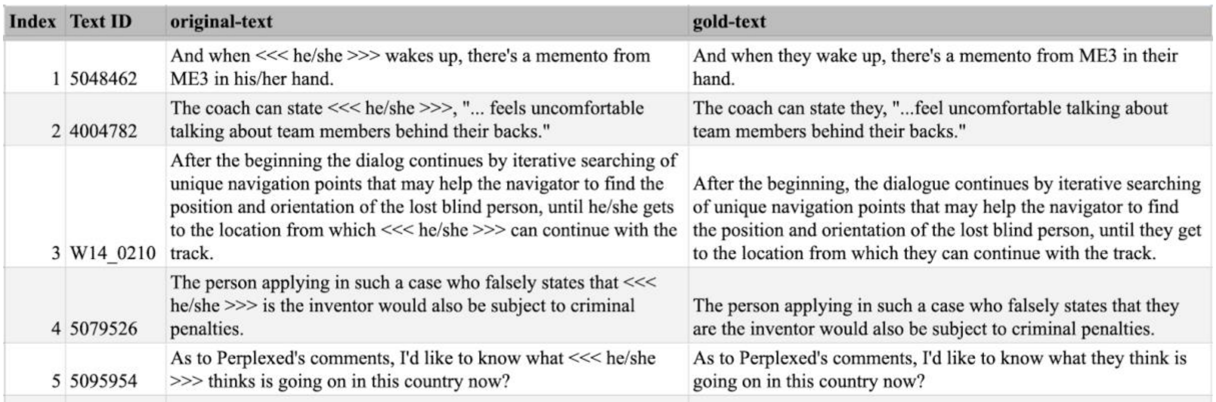
For the creation of the dataset, no automated preprocessing was involved. Each sentence was reviewed by the researcher for preprocessing and cleaning. The cleaning involved the correction of spacing and spelling of the raw data since the data from the CPQ webpage has different tokenization processes. The original data for each genre from CQP Web can be found online.[3]

Due to the format of text from the CQP webpage, the token *that* appears as *that/IN*.

[1] https://gucorpling.org/cqp/

[2] 573 was an arbitrary number chosen due to the expectation of what the researcher can do within the given time frame.

[3] https://github.com/jessicacusi/singular-they/tree/d5e19a5fd099e22234a05ffc101c3eec8e2ba434/raw_data



Index	Text ID	original-text	gold-text
1	5048462	And when <<< he/she >>> wakes up, there's a memento from ME3 in his/her hand.	And when they wake up, there's a memento from ME3 in their hand.
2	4004782	The coach can state <<< he/she >>>, "... feels uncomfortable talking about team members behind their backs."	The coach can state they, "...feel uncomfortable talking about team members behind their backs."
3	W14_0210	After the beginning the dialog continues by iterative searching of unique navigation points that may help the navigator to find the position and orientation of the lost blind person, until he/she gets to the location from which <<< he/she >>> can continue with the track.	After the beginning, the dialogue continues by iterative searching of unique navigation points that may help the navigator to find the position and orientation of the lost blind person, until they get to the location from which they can continue with the track.
4	5079526	The person applying in such a case who falsely states that <<< he/she >>> is the inventor would also be subject to criminal penalties.	The person applying in such a case who falsely states that they are the inventor would also be subject to criminal penalties.
5	5095954	As to Perplexed's comments, I'd like to know what <<< he/she >>> thinks is going on in this country now?	As to Perplexed's comments, I'd like to know what they think is going on in this country now?

Figure 2. Example of data in ST dataset.

Therefore, all instances of *that/IN* in both the original and gold sentences are changed to *that*. The original sentences also contain the <<< >>> brackets that encase the query search term "he/she." For the initial development of the rule-based baseline model, the brackets were kept. However, for future use, potential users of this dataset can easily remove the arrows for "cleaner" data.

3.3 Uses and Distribution

The ST dataset has been used for one task thus far, which was to test the rule-based model (discussed in the next section). 20% of the data (86 sentences) was used for testing; the remaining 80% (343 sentences) can be used for future training of different models.

There are no harms or ethical concerns with the contents of the dataset, as no personally identifiable information is shared. All the data, since it comes from the ACL Anthology and collections of COCA, is commonly and publicly used and shared.

This dataset was created as a starting point for developers to train ML models with gender inclusive data. Because the dataset contains raw and gold-labeled sentences, it could be used for a variety of tasks. All raw data and the final ST data could be found online.[4]

4 Rule-Based Model

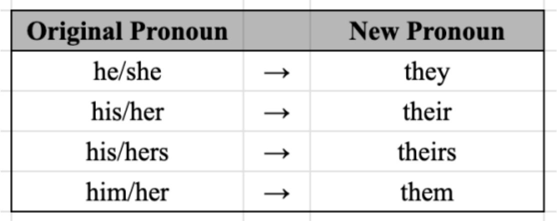
The RB model using the ST dataset was developed using Python (v. 3.7). It is published alongside the dataset on GitHub as the *rb_model.py*.[5]

4.1 Preprocessing

In contrast to the creation of the dataset itself, the RB model has a preprocessing step. This is done to make the original sentence and gold standard sentence as identical as possible before comparison. Preprocessing includes lowercasing and the removal of double spacing and the arrows (<<< >>>) that encase "he/she."

4.2 Pronoun Replacement

After preprocessing, all variations of binary gender pronouns are replaced with the corresponding SGT form using *re*'s substitution method. The pronoun changes are shown in Figure 3.



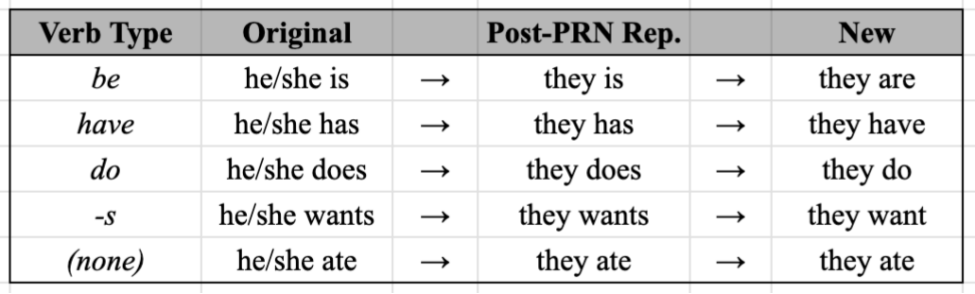
Original Pronoun		New Pronoun
he/she	→	they
his/her	→	their
his/hers	→	theirs
him/her	→	them

Figure 3. Pronoun changes in RB model.

As previously mentioned, during the data collection stage, sentences were retrieved specifically for the query *[word = "he/she"]*. However, if the selected sentence contained any other binary pronoun form listed under the "Original Pronoun" column in Figure 3, the pronoun was changed based on the model's rules.

[4] https://github.com/jessicacusi/singular-they

[5] https://github.com/jessicacusi/singular-they/blob/d5e19a5fd099e22234a05ffc101c3eec8e2ba434/rb_model.py



Verb Type	Original		Post-PRN Rep.		New
be	he/she is	→	they is	→	they are
have	he/she has	→	they has	→	they have
do	he/she does	→	they does	→	they do
-s	he/she wants	→	they wants	→	they want
(none)	he/she ate	→	they ate	→	they ate

Figure 4. Sequence of changes made in RB model.

4.3 Verb Adjustment

After the pronouns are changed to SGT forms, the returned data goes through verb adjustment. There are 4 possible adjustments: *be* verbs, *have* verbs, *do* verbs, and verbs ending with *-s* (Figure 4). For the first three options, the rules are simple. Because the verb has a specific third-person singular and plural form, the verb has one-to-one replacement after the pronoun is changed. For the verbs that end with -s, a rule was created to detect if a word following “they” ended with -s. If it did, it was replaced with all characters of that word except for the final *-s*, resulting in the third-person singular form. For example, as shown in Figure 4, “he/she wants” transforms to “they wants” before ending as “they want.” Lastly, if the model does not find a verb of any of the above categories, the verb remains the same.

4.4 Evaluation

After all of the original, raw sentences have been corrected, the RB model was evaluated for accuracy. To obtain this metric, each corrected sentence was considered correct (i.e., given a score of 1) if it was completely identical to the gold standard sentence from the ST dataset. Accuracy was obtained by taking the sum of the scores and dividing by the total number of sentences. Based on these methods, the RB model received an accuracy of ~0.465.

5 Discussion

There are 3 main limitations with the ST dataset and the corresponding RB model: the query search, dataset size, and factors within the rule-based model itself.

Firstly, the query search was intentionally limited to specific instances of “he/she” for simplicity. Other forms of the binary gender forms such as “he or she” or “(s)he” were left out. Therefore, the initial data collection was biased to only contain “he/she.” Furthermore, instances of the generic masculine “he” and generic feminine “she” were also excluded. These instances are much more difficult to retrieve based solely on a query search without enough relevant context.

Secondly, the dataset is relatively very small. Although the initial goal was roughly 573 sentences, unexpected issues with the data and time constraints decreased the final total to 429. As mentioned previously, 116 sentences from the COCA TV & Movies genre were deleted because the data was too messy. This is likely due to the nature of this specific corpus data since it contains very colloquial language from subtitles.[6] Additional deletions to the final dataset were made during the validation stage. The limited size is also due to the mere expensiveness of human data collection and validation.

Lastly, the rule-based model may be limited in its development and applications. There are only a handful of designated rules (i.e., *be* verbs, *have* verbs, *do* verbs, regular verbs ending with *-s*, and none of the prior). These rules were chosen based on frequency of the verbs and understanding of language patterns, but there is a high chance that additional rules were missed. If rules were excluded, “incorrect” raw data would remain the same and not have the SGT. The gold-standard sentence would also be different, thus reducing accuracy. Furthermore, the RB model was developed specifically for the “singular they” dataset; it has not been tested on

[6] `https://www.english-corpora.org/coca/help/coca2020_overview.pdf`

other data. Overall, the model is best to be considered as a baseline for comparison to future, more advanced models, such as one built with a Transformer architecture.

6 Conclusion

The goal of this project was to increase gender inclusivity in datasets by removing instances of binary gender pronouns and replacing them with the singular generic "they." A gold-standard dataset consisting of 429 was created and used to test a rule-based model that changed pronouns and verbs according to four categories: *be*, *have*, and *do* verbs and verbs ending in *-s*. Overall, the rule-based model did not receive high accuracy, likely due to the inconsistencies with the raw data (e.g., spacing, spelling) that make exact matches difficult. Regardless, this project can be a productive first step to prevent future models from using any gender exclusive and sexist language.

6.1 Future Work

Due to time constraints for this project, there are different avenues that were unable to be tackled in more depth. Thus, there are many future directions for this project.

As with any computational or research project, the more data, the better. Thus, future students or researchers can preprocess and gold-label additional pre-collected raw data. Alternatively, future researchers can gather new, more recent data to add to the ST dataset. For the scope of this project, all the data comes from English. If binary gender pronouns are used as often in other languages, future researchers could look into a similar project for those languages. That being said, American English culture already has some divided perspectives, so research should be done beforehand to address the societal validity of the corrections.

In regard to the computational aspect, the current RB model can be applied to different datasets in order to achieve a quick solution to transition to singular generic "they" language. Future researchers can also build a more complex model using Transformer architecture so that the SGT corrections are not limited to the pre-defined rules.

Acknowledgements

The author would like to thank Katelyn Carroll for validating the "Singular They" Dataset and for the helpful comments she provided, as well as Professor Shabnam Tafreshi for guidance and support for the project.

References

APA Style. (July 2022). Singular "They". https://apastyle.apa.org/style-grammar-guidelines/grammar/singular-they#:~:text=The%20singular%20%E2%80%9Cthey%E2%80%9D%20is%20a,avoid%20making%20assumptions%20about%20gender

Balhorn, M. (June 2004). The Rise of Epicene *They*. *Journal of English Linguistics (32)*. DOI: 10.1177/0075424204265824

Baron, D. (2018, September 4). *Oxford English Dictionary*. Oxford English Dictionary. https://public.oed.com/blog/a-brief-history-of-singular-they/

Baumler, C. & Rudinger, R. (July 2022). Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3426–3432.

Bradley, E.D. (January 2020). The influence of linguistic and social attitudes on grammaticality judgments of singular 'they'. *Language Sciences 78*. https://doi.org/10.1016/j.langsci.2020.101272

Doll, Jen. (January 2013). *The Singular 'They' Must Be Stopped. https://www.theatlantic.com/culture/archive/2013/01/singular-their-affront-good-writing/319329/*

Foertsch, J. & Gernsbacher, M.A. (March 1997). In search of gender neutrality: Is Singular *They* a Cognitively Efficient Substitute for Generic He? *American Psychological Society 8(2)*.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K. (November 2021). Datasheets for Datasets. *Communications of Association for Computing Machinery*, *64*(12), 86-92. https://doi.org/10.1145/3458723

LaScotte, D.K. (February 2016). Singular They: An Empirical Study of Generic Pronoun Use. American Speech 90(1). DOI: 10.1215/00031283-3509469

Mackay, D.G. (December 1980) On the Goals, Principles, and Procedures for Prescriptive Grammar: Singular They. Language in Society 9(3), 349-367. https://www.jstor.org/stable/4167168

Mackay, D.G. & Fulkerson, D.C. (1979). On the Comprehension and Production of Pronouns. *Journal of Verbal Learning and Verbal Behavior*, 18, 661-673.

Martyna, W. (1980). Beyond the “He/Man” Approach: The Case for Nonsexist Language. *Signs*, *5*(3), 482–493. http://www.jstor.org/stable/3173588

Maryskova, A. (2021). The Use of Singular They vs. Gendered Pronouns. Masaryk University Faculty of Arts: Department of English and American Studies.

Miller, M. M., & James, L. E. (2009). Is the generic pronoun he still comprehended as excluding women? *The American Journal of Psychology*, *122*(4), 483–496. http://www.jstor.org/stable/27784423

Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B. (2018). Gender Bias in Coreference Resolution. *Proceedings of NAACL-HLT 2018, 8-14*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Change, K-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of NAACL-HLT 2018*, *15–20*.