LING807 Final Paper- Virginia Uhi

A Test on spaCy's Coreference Resolver with Regards to Pronoun 'They' Counterparts

Introduction

In modern day, more individuals started to identify themselves with gender-neutral pronouns such as ‘they/them’. Merriam-Webster officially added the non-binary pronoun "they" as an entry in 2019, reflecting its growing usage and acceptance in English (*Singular “They,”* 2019). This has also been a hot topic in computational linguistics where language models will have to adapt to this gradual change. Current literature tested major models such as GPT, OPT and PaLM for accuracy and data bias in gender labelling. This paper will examine spaCy, an accessible NLP system, with similar tests on a smaller scale. Mainly to investigate the accuracy of spaCy coreference in identifying singular and plural ‘they’ respectively. Lastly, this paper will discuss the system’s weakness and how it can be improved.

Literature Review

Past Literature on Gender in Computational Linguistics

Large Language Models (LLMs) should reflect current human language phenomena and gender-neutral pronouns is one of the current hot topics for computational linguistics. Pronouns have evolved ‘from a closed class of words with few members to a much more open set of terms to reflect identities (Lauscher, Crowley, & Hovy, 2022). Various tests have been conducted continuously to evaluate LLMs whether they are accurately reflecting human speech. Unfortunately, results have been undesirable for English language models.

Hossain, Dev, and Singh (2023) tested common English LLMs (BART, T5, GPT-2, GPT-J, OPT and BLOOM) at that time according to the MISGENDERED framework they developed for the research. First, they manually compiled templates with the pronoun position blank. They then

populate the templates with controlled data to try to mitigate the effect of biased data training. Finally, a prediction test is given to the LLMs, they are prompted to fill in the templates on their own. They undergo constrained decoding and produce an output of the test to fill in the blank. The results show that the accuracy rate for predicting the correct binary or non-binary pronouns is low. Meanwhile, Brandl, Cui, and Søgaard (2022) research on an eye-tracking study for Swedish speakers shows that they do not have extra difficulty in comprehending gender-neutral pronouns compared to regular binary pronouns. The study proposed that Swedish is linguistically structured to better accommodate gender-neutral pronoun 'hen', thus Swedish speakers process binary and non-binary pronouns similarly.

Various literature discusses potential problems, including LLMs that are not updated once they are published. But the biggest reason that past literatures are pointing towards is that English LLMs are built and trained on biased data, especially gender biased data. 'Newly developed algorithms do not test their models for bias with limited definitions of gender bias and lacking evaluation baselines and pipelines.' (Stanczak & Augenstein, 2021). The industry does not have a complete evaluation framework to safeguard the balance of datasets for machine learning, it is understandable that developers rarely test their newly developed models.

More importantly, gender bias in co-reference is highly disruptive in an inclusive environment. It is identified that rule-based systems are the most biased, while feature-rich systems are the least biased (Zhao, Wang, Yatskar, Ordonez, & Chang, 2018). Ruled-based model is an older approach by hardwiring rules to generate a suitable output. It does not account for exceptions and cannot progress or evolve with time without manual updates. Contrastingly, a feature-rich approach does not require hardcoding, but a probability system. The more features are matched, the more likely a certain output is correct. This allows more flexibility in adapting to new knowledge, in this case,

gender-neutral pronouns. The machine has more space to learn to adjust to new features added. In coreference, it is important to be flexible in modern times, since personal pronouns are based on preferences and not rules.

The Three Types of 'They'

'They', 'them', 'their(s)' and 'themselves' are third-person plural pronouns in nominative, objective, genitive and reflexive cases respectively. (Quinn, 2005) Although 'them*self*' is also a variation of 'they', it is commonly used to refer to singular nouns, in contrast to 'them*selves*'. According to the Merriam-Webster Dictionary, the word is defined as a) 'Those ones; those people, animals, or things in plural'; b) 'Refer to people in a general way or to a group of people who are not specified.' and lastly c) 'Used with a singular indefinite pronoun antecedent.' ("Definition of THEY," 2016).

Plural 'they' are commonly used when referring to multiple entities and has been a norm in English for years. For the first definition, (1) - (4) have multiple subjects in the sentence, thus a plural pronoun indicating the number of subjects is more than one.

(1) **Chris** and **Alex** are siblings, ***they*** love each other.

(2) **Tom** and **Jerry** are good people, you can trust **them.**

(3) **Their** car is in parked outside, call either **Peter** or **John** to move it.

(4) The book is **theirs**, the **twins** left it here.

However, the usage of the singular general 'they' can be dated back to the 1930s but did not become popular until modern time English. According to Fries (1940), Geoffrey Chaucer did not use 'them' as the objective and 'their' as the genitive form of 'they' in English. Instead, he used

‘hem’ for the objective form and ‘her’ or ‘hir’ for the genitive form. This shows that even in plural, there were still gender elements incorporated into the pronouns. In modern times, compared to Spanish or German, English has developed into a less gender-dependent language. Hence gendered plural pronouns such as ‘hem’ and ‘hir’ have been abandoned and replaced. Indefinite pronouns are great examples in demonstrating the usage of singular ‘they’ to refer to a general group. Other than the history of plural pronouns, Fries (1940) also suggested that ‘the use of number forms in secondary words that has emerged in the development of English is a concord based primarily on the number idea emphasized in the primary word rather than on its form.’ Meaning that it is the idea of plurality determines the form of pronoun rather than the indefinite pronouns itself. In (5), ‘Some*one*’ is essentially referring to one person, but the idea of that ‘some*one*’ can be multiple different persons, so it proceeds to take a plural reference pronoun.

This idea extends beyond indefinite pronouns and into nouns that possess the possibility of referring to more than one entity. It mainly tries to form a common ground or idea for situations. (6) illustrates the speaker trying to create a common topic between the listener and themself. The context of persuading an individual to rejoin the team makes ‘the company’ an idea of plurality. Even though the antecedent is in singular form on the surface, it holds to concept of multiplicity which ultimately determined the plural form of the reference pronoun.

(5) **Someone** left **their** keys on the table.

(6) **The company** wants you, **they** need you.

English lacks epicene pronouns, thus needing to borrow neutral pronouns from plural pronouns, that can accommodate both male and female subjects (Baron, 1981). This type of singular personal ‘they’ is used to eliminate gender bias in situations, hinting that the antecedent can be male, or

female as seen in (7). It also gave rise to individuals self-identifying with gender-neutral pronouns in getting rid of gender labels in modern societies. In hopes of eliminating the power dynamic with genders in language, more English users have accepted and incorporated singular personal 'they' like (8), into their lexicon.

(7) **The doctor** is coming, **they** will be here soon.

(8) **Jodie** is an award-winning chef. **They** own 10 restaurants across the world.

Psycholinguistics on Gender

Gender is perceived differently by individuals. According to Tripp and Munson (2021), social background and knowledge, demographic, ideology and language variations all affect gender perception. In other words, there is not a strict framework that can map out the general process of gender perceiving. It is often associated with one's belief in others of what they can and cannot become, linking their judgement to desirability. For example, Conservatives who believes in only binary gender, Gregory Thomson will be perceived as a 'man' because that name belongs to a male and only male. This phenomenon is called metaphysicality, where one's physical appearance or features are assumed based on meta information that does not concretely exist.

When it comes to gendered languages, instead of having variety in interpretation of gender, default genders can help keep track of referents in discourses by disambiguating referential construction (Berkum, 1996). This phenomenon is called gender priming, gendered nouns and pronouns work together to prepare speakers to grasp the discourse quicker. The research investigated the response and processing time of Dutch speakers - gendered language speakers, in multi-participants discourse. The results show that Dutch speakers process complex discourse similarly with simple discourse. There is no concrete evidence showing that gender priming facilitated the processing

time of speakers, but the author also argued that there is no evidence to disprove gender priming either. Looking back on English, is there another way it can perform similar effect such as gender priming as a non-gendered language. Current literatures on gender perception in English has many social contexts that machines cannot account for, especially when they lack world knowledge. At this point, there is still no discovery in English linguistics structure or features that allow LLMs to identify genders correctly in different social context.

Coreference Evaluation Metrics

The oldest major coreference metrics is the algorithm for Message Understanding Conference (MUC). It is widely discussed by Vilain, Burger, Aberdeen, Connolly, and Hirschman (1995), with continuous improvements in the coming years. This metrics counts the common links between the 'truth' (manual annotations) and system output. The precision score is calculated by the common links between the truth and system output divided by the total output links. The recall score is calculated by the common links between the truth and system output divided by the total truth links. A fatal flaw in the MUC is that the metric favors systems with less output since the smaller the denominator, the higher the score is more likely to be regardless of the number of common links.

$$\text{Precision score} = \frac{Common\ links}{Total\ output\ links}$$

$$\text{Recall score} = \frac{Common\ links}{Total\ truth\ links}$$

Fig. 1. Visual representations of the precision and recall score equations.

B-Cube proposed by Bagga and Baldwin (1998) tries to overcome MUC bias by calculating recall and precision for each entity instead of calculating the number of links between all entities in the

document. B-cube successfully eliminated quantity bias but a new problem of double counting the same entity arises. Since a precision and recall score is calculated for each entity independently, the metric overlooks the relationship between a single entity, and it should not overlap during calculations. The makes the final evaluation score unreliable.

Luo (2005) proposed another metric called Constrained Entity-Aligned F-Measure (CEAF) that aims to measure the quality of system over quantity. CEAF aligns each output with at most one 'truth', and vice versa. The metric calculates precision and recall score based on the number of correct alignments. The scores are then adjusted based on system output size to avoid quantity bias. Mention-based CEAF reflects the percentage of correctly mentions in the entities. Entity-based CEAF reflects the percentage of correctly recognized entities. The problem with CEAF scores is that the alignment step serves like a screening step, it excludes the correct system output under the wrong entities. This results in a loss of accuracy in the calculation since not every correct mention or output is accounted for.

The latest evaluation metric, the Link-based Entity-Aware metric (LEA), is proposed by Nafise Sadat Moosavi and Strube (2016) to overcome all past shortcomings. Its model is like the classic MUC, but it adjusted its linking calculation to avoid output size bias. This allows LEA to perform one-to-many mapping for entities contrasting to B-cube, where the same entities are calculated repeatedly. It also accounts for all correct links since it does not have an alignment step like CEAF. In general, LEA design is better since it accounts for all the links and effectively eliminated quantity bias.

Methodology

What is spaCy and Why SpaCy

'spaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. It's designed specifically for production use and helps you build applications that process and "understand" large volumes of text.' ("Facts & Figures · SpaCy Usage Documentation," n.d.). It is an accessible and customizable database for NLP systems. Despite its initial release in 2015, developers have been actively working on spaCy to improve the system. This paper aims to investigate a new, emerging way of gender pronoun usage, spaCy as a well-maintained platform will be a great site to see if the system is up to date with current language changes. Its flexibility and compatibility with Python make it a perfect small-scale platform to replicate tests that were performed at a larger scale in past literature.

Experiment

This experiment mainly investigates the accuracy rate of spaCy coreference on singular and plural 'they' counterparts (they/them/their/theirs/themself/themselves). The spaCy small English copra is used with an experimental spaCy extension component called Coreference Resolver. Six articles are analyzed, three regular news articles and three news articles that report extensively on gender-neutral pronouns. Regular articles are controlled data that provides a baseline for the accuracy rate of spaCy. Those articles contain mostly plural 'they' and seldom singular 'they', the other three queer articles are vice versa. All six articles are manually annotated and processed through spaCy, then the results are compared to each other to calculate the precision and recall score.

Two-step Evaluation

Basic coreference evaluation can take part in two steps: (i) Determining mentions, and (2) deciding whether mentions are coreferent or not (Cai & Strube, 2010). Ideally, the two steps should also be scored independently of each other (Pradhan et al., 2014). Step one, all mentions of potential referents from a document. Step two, determining whether the identified referents refer to the same and correct entity. A two-step evaluation is required for the analysis. Step one, due to the scope and focus of the experiment, only referents of potential 'they' counterparts are labelled in the manual annotation. For example, gendered pronouns (he/she counterparts) are not labelled. Step two, all labelled referents are then determined whether they refer to the correct entity.

For mention detection, the recall and precision score are calculated according to the Confusion Matrix. Recall score is calculated by True positives divided by total positives, while precision score is calculated by true positives divided by the sum of true positives and false negatives. Data is recorded on a spreadsheet, identified tokens were labelled as '1' under its designated columns, and '0' for the opposite. True positives are tokens with '1' in both manual and spaCy column; false positives are tokens with '0' in manual and '1' in spaCy; True negatives are tokens with '0' for both manual and spaCy; False negatives are tokens with '1' in manual and '0' in spaCy.

System output	Truth (Manual annotation)		
Positive	True positive	False positive	Recall = $\frac{tp}{tp + fp}$
Negative	False negative	True negative	
	Precision = $\frac{tp}{tp + fn}$		

Fig. 2. Visual representations of the Confusion Matrix

For step two, none of the mentioned coreference metrics is used in this small-scale experiment. Instead, labelled referents are separated into heads and non-heads. Then, non-heads are further analyzed in addition with their pronoun properties. For example, which type, singular or plural ‘they’ the mention belongs to. A matching rate is used to determine how accurate spaCy can label referents correctly compared to the ‘truth’.

Hypothesis

SpaCy is predicted to produce similar results as past literatures despite it being a smaller and more updated model. The score for mentions identification will be high for both regular and queer articles since past research shows little confusion in identifying nouns and entities across all languages. However, queer articles will have a significantly lower score for coreference evaluation due to the lack of context knowledge compared to humans. Plural ‘they’ will also significantly outperform singular ‘they’ due to its popularity and history in the English language.

Results and Discussion

Data Results

The average of precision and recall scores for regular and queer articles are above 80%, with no significant differences between individual article scores. These results are in line with past literature and findings, that machines can identify referents, especially pronouns which take up most of the referent list.

	Precision	Recall	Average Precision	Average Recall
Text 1	83%	81%	89%	86%
Text 2	85%	76%		
Text 3	100%	100%		
Text 4	76%	76%	80%	82%
Text 5	75%	90%		
Text 6	90%	81%		

Fig. 3. Mention scores for all text

Moreover, almost all ‘they’ counterparts were identified, with only 1-2 misses. The rest of the errors mostly happen with chain heads, or antecedents' identification. SpaCy sometimes picks up false referents, usually movie or show titles. It is inferred that since these titles are not under spaCy’s ‘entity’ library, they are treated as a regular noun. For example, in figure 4, Killers of the Flower Moon is considered as a title since the first Alphabet is capitalized for each word, a common indicator for proper nouns such as names. However, the system selects ‘Killers’, a part of the movie title as the head. This shows that the system considers the title as five separate entities instead of one.

Killers of the Flower Moon star Lily Gladstone has opened up about how using she/they pronouns is connected to the performer’s Indigenous background and “partly a way of decolonizing gender.”

Fig 4. Excerpt of Text 4. The yellow highlight represents manual annotation reference chain, the red text represents spaCy output chain.

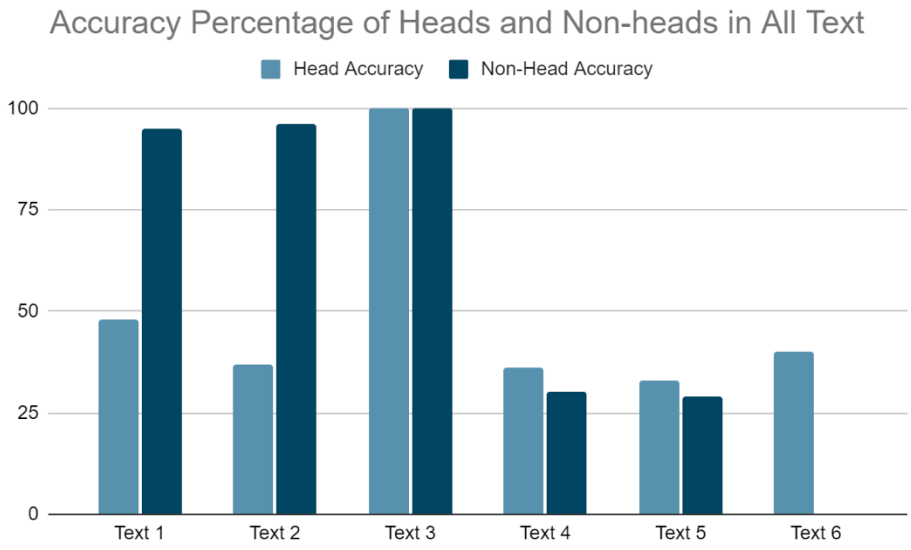

Fig. 5. Accuracy Percentage of Heads and Non-heads in All Texts

As shown in Figure 5, spaCy produces false labels (errors) in heads rather than non-heads. Meaning that spaCy has difficulty in correctly identifying antecedents since it requires contextual knowledge. All scores are high for Text 3, it is an article about humans and dogs. The text only had three main entities: the narrator, the dogs and a female friend. These three entities have their own set of distinct pronouns. The narrator, and his friend never appear in the same environment, and the dogs were always described in packs. Thus, making it clear for spaCy to generalize male pronouns belong to the narrator, female pronouns belong to the friend, and plural pronouns to animals.

The coreference score for step two has significantly difference between regular and queer articles. The average coreference score for non-head in regular text is 97% while queer text only has 20%. These findings are also in line with previous literature that machines have much difficulty in recognizing singular ‘they’. This phenomenon can be first explained by the frequency of singular

and plural ‘they’ between the two types of articles. In modern day, self-identifying as ‘they/them’ pronouns are still emerging. Regular articles have yet to incorporate such expressions into their writing. However, in queer articles, where the goal is to raise awareness for such a matter, an extensive use of these difficult pronouns is used. Although spaCy has difficulty in processing all singular ‘they’, the significant decrease in performance rate might be caused by queer articles having a substantial amount of singular ‘they’ than regular articles.

Weaknesses of SpaCy

SpaCy showcased its strength on tokenizing and categorizing text to facilitate better language processing. However, during the experiments, it is discovered that spaCy only has plural ‘they’ in its POS tagging and morphology list. Every entry for ‘they’ counterparts is identical despite being in different syntactic positions or context. This suggests that all 'they' were set to plural by default, and the system will not be able to recognize singular ‘they’ because it was never an option to do so. These findings echos the issue of data bias, where the library itself is biased by denying the usage of singular ‘they’.

They: PRON PRP nsubj Case=Nom|**Number=Plur**|Person=3|PronType=Prs

Them: PRON PRP pobj Case=Acc|**Number=Plur**|Person=3|PronType=Prs

Their(s): PRON PRP$ poss **Number=Plur**|Person=3|Poss=Yes|PronType=Prs

Theirselves: PRON PRP dobj Case=Acc|**Number=Plur**|Person=3|PronType=Prs|Reflex=Yes

Fig 6. Entries from spaCy token.morph for ‘they’ counterparts

Another weakness of SpaCy is that it has difficulties with abstract ideas. In other words, it is hard for the system to determine whether the abstract object should be a referent or not. Incorrectly

identifying an abstract object, especially temporal related nouns as an antecedent is a common mistake in SpaCy. For instance, ‘future’, ‘time’ and ‘Wednesday’ were labeled as the heads of ‘they’. Any token within a text has a possibility to become a referent. However, in context, they are nouns that happened to be in between the intended head and non-head. Reasons for this mistake are unclear, but the repeated errors throughout the test points towards a fundamental error in the system design.

Possible patterns on how SpaCy labels

It is speculated relevancy may play a role in coreference labelling, as suggested by Berkum (1996). Gender recency is one of the main explanations for Dutch speakers can process complex discourse similarly to regular discourse. The closer the gender marker is to the referee, the shorter time it takes to pinpoint and process the discourse. This statement is also supported by Tai, Socher, and Manning (2015), people are more likely to retrieve information from a more recent discourse marker due to our short-term memory based on syntactic relevancy. In recent years, researchers have started to doubt such findings and argue that syntactic relevancy only performs well on a sentence level and short discourses, its effect might be limited in larger environments (Jayaswal, 2020).

In spaCy, a certain degree of recency is observed during the test. When multiple plural nouns are in a sentence, the system will choose the closest one such as in (9) and (10). If there is another plural noun intervened with the chain, the system will start a new chain with the closer pronouns. In (11) and (12), although semantically the pronouns are still referring to the previous antecedent, the head was switched to another plural noun within the paragraph.

(9) Rightly or not, that will combine with the many variables over which politicians have more control, to help determine Canadians' mindset about carbon pricing when they send a message to the rest of the world in the fall.

(10) And there were lots of women historically and still now who are given men's names. They fulfill more of a man's role in society as far as being provider, warrior, those sort of things.

(11) Still, McNeely said disease can complicate a relationship. When she and her boyfriend got a cat together, McNeely said they had to consider whether he could take care of the pet without her. When they discuss the prospect of marriage, she worries about whether debts* related to her illness would transfer to him after she dies.

(12) Liberal-aligned groups may respond by painting Mr. Scheer as a threat to the planet's future. They could respectively have financial backing from corners of the resource sector and from environmental groups, or those interests could go it alone with their own third-party ads.

**debts* is a cataphora, acting as the head of the previous *they*.

Fig. 6. Excerpt of Text 1, 2 and 4. The yellow highlight represents manual annotation reference chain, the red text represents spaCy output chain.

Other than recency, another observable pattern is that the machine priorities nouns with overt plural morphology as antecedents over those without. Nouns that are labelled 'Number=Plur' or two nouns directly connected by 'and' are always linked to 'they' pronouns. For example, (13) and (14) show that spaCy identifies overt plurals over multiple singular entities. These results suggest that the system relies heavily on POS tagging to perform coreference tasks as the two examples show no recency on syntax nor sentence level.

(13) After weeks of online flirting, Patrick Bardos was en route to meet Anne Marie Cerato for their first date at a coffee shop in downtown Toronto.

(14) He said it's not uncommon for people to sever ties, even marriages, with partners rather than confront the prospect of losing a loved one to cancer, and by proxy, face their own mortality. But while some couples collapse under the strain of sickness, Rutledge said, for others, it can heighten emotional connections.

Fig. 7 Except of Text 1. The yellow highlight represents manual annotation reference chain, the red text represents spaCy output chain.

Ultimately, no concrete evidence is found on the pattern or rule of how spaCy's coreference resolver makes the chains, above discussions are based on limited data and trends.

Improvements for SpaCy

Throughout the test, an inconsistent of correct labelling of singular 'they' is found (8 total Chains). 5 Chains were singular general 'they', headed by words like 'people' and 'someone'. These are typical usage for singular general 'they'. The other 3 chains were correct identification of gender-neutral pronouns. All chains do not share any observable similarities, thus it is hard to determine if the system is doing something correctly or is by pure luck. However, an update on spaCy library, especially 'they' counterparts is needed to eliminate gender bias. Datasets should not be set to only one default setting when there is another type of usage of the word. The library should start to adapt to the new type of 'they' by incorporating it into its POS tagging property. It requires semantic knowledge to determine when the pronoun is referring to a singular antecedent, however, spaCy seems to be learning common usage of singular 'they' such as in referring to a general group already. Added more options to its part of speech tag might allow space of growth under that tag as more data is provided.

Conclusion

With gender-neutral pronouns starting to emerge, the need for LLMs to accommodate singular ’they’ is increasing. SpaCy is the biggest available NLP platform, and the system is worth testing and improved on. Although its Coreference Resolver is still in its experimental stage, it can identify most of the referents compared to manual annotations. However, the accuracy rate for correctly labelling entities remains low, and especially low for singular ’they’. The machine is set to recognize ’they’ counterpart pronouns as plural indicators, illustrating a fundamental cause for difficulty in pairing them to singular nouns. The results of this paper may be limited since the tests were done on spaCy‘s smallest English corpora, more tests should be performed on the medium and large corpus to better locate patterns. Sample size of this test is also limited, not all topics of articles were analyzed. Future studies may consider if article topics and genre influence coreference accuracy score. Furthermore, this experiment should be replicated using a major coreference evaluation metrics for more in depth results.

References

Bagga, A., & Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. *Proceedings of the Linguistic Coreference Workshop at the First International Conference on Language Resources and Evaluation (LREC’98)*.

Baron, D. E. (1981). The Epicene Pronoun: The Word That Failed. *American Speech*, *56*(2), 83. https://doi.org/10.2307/455007

Berkum, V. (1996). *The psycholinguistics of grammatical gender : studies in language comprehension and production.* Nijmegen: Nijmegen University Press.

Brandl, S., Cui, R., & Søgaard, A. (2022). How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2204.10281

Cai, J., & Strube, M. (2010). *Evaluation Metrics For End-to-End Coreference Resolution Systems*. 28–36.

Definition of THEY. (2016). Retrieved from Merriam-webster.com website: https://www.merriam-webster.com/dictionary/they

Dhingra, B., Jin, Q., Yang, Z., Cohen, W. W., & Ruslan Salakhutdinov. (2018). Neural Models for Reasoning over Multiple Mentions Using Coreference. *ArXiv (Cornell University)*. https://doi.org/10.18653/v1/n18-2007

Facts & Figures · spaCy Usage Documentation. (n.d.). Retrieved from Facts & Figures website: https://spacy.io/usage/facts-figures

Fries, C. C. (1940). *American English Grammar*. Irvington Publishers.

Hossain, T., Dev, S., & Singh, S. (2023, July 1). MISGENDERED: Limits of Large Language Models in Understanding Pronouns (A. Rogers, J. Boyd-Graber, & N. Okazaki, Eds.). https://doi.org/10.18653/v1/2023.acl-long.293

Jayaswal, V. (2020, September 15). Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score. Retrieved from Medium website: https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262

Lagunoff, R. (1997). *Singular They*.

Lauscher, A., Crowley, A., & Hovy, D. (2022). Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2202.11923

Luo, X. (2005). On coreference resolution performance metrics. *Empirical Methods in Natural Language Processing*. https://doi.org/10.3115/1220575.1220579

Nafise Sadat Moosavi, & Strube, M. (2016). Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. *White Rose Research Online (University of Leeds, the University of Sheffield, University of York)*. https://doi.org/10.18653/v1/p16-1060

Pradhan, S., Luo, X., Marta Vilar Recasens, Eduard Hovy, Ng, V., & Strube, M. J. (2014). Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. *Meeting of the Association for Computational Linguistics*, 30–35. https://doi.org/10.3115/v1/p14-2006

Quinn, H. (2005). *The distribution of pronoun case forms in English*. Amsterdam ; Philadelphia: John Benjamins Pub.

Singular "They." (2019, September). Retrieved from www.merriam-webster.com website: https://www.merriam-webster.com/wordplay/singular-nonbinary-they

Stanczak, K., & Augenstein, I. (2021). *A Survey on Gender Bias in Natural Language Processing*. https://doi.org/10.48550/arxiv.2112.14168

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *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)*. https://doi.org/10.3115/v1/p15-1150

Tripp, A., & Munson, B. (2021). Perceiving gender while perceiving language: Integrating psycholinguistics and gender theory. *WIREs Cognitive Science*, *13*(2). https://doi.org/10.1002/wcs.1583

Vilain, M., Burger, J. D., Aberdeen, J. S., Connolly, D., & Hirschman, L. (1995). *A model-theoretic coreference scoring scheme*. https://doi.org/10.3115/1072399.1072405

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. https://doi.org/10.18653/v1/n18-2003

Appendix (In GitHub repository)

https://github.com/GinnyChan27/LING807_Final_Prj/tree/main

*Experiment codes can also be found on the repository

Appendix A: Text 1 Manual Annotation and SpaCy Annotation Records

Appendix B: Text 2 Manual Annotation and SpaCy Annotation Records

Appendix C: Text 3 Manual Annotation and SpaCy Annotation Records

Appendix D: Text 4 Manual Annotation and SpaCy Annotation Records

Appendix E: Text 5 Manual Annotation and SpaCy Annotation Records

Appendix F: Text 6 Manual Annotation and SpaCy Annotation Records

Appendix G: Evaluation Spreadsheet