COSC420 Assignment 2 - Josh Whitney (4442561)

Introduction

In this assignment, I was tasked to examine the issue of training bias set bias in language models, focusing on gender bias. For the purposes of this project, I worked with the gpt2-simple model, using their provided Google Collaboratory notebook to fine-tune my own models, using the provided Brown corpus (the tagged version), or variants of it as a training set. I set about this task be comparing a gpt2-simple model fine-tuned on the first 50% of the Brown corpus, to a gpt2-simple model fine-tuned on a modified version of the same text (a manipulated corpus)- in which I attempted to swap out binarily-gendered words for gender-neutral terms, with the aim of reducing various forms of gender bias in the generated outputs.

Section 1: Corpus Manipulation

To manipulate the Brown corpus, I first cut it down to roughly half the original size by cutting the text file into two separate text files, of the 57,076 total lines of text, and placing the first 28,533 lines of text in one file, and the rest in the other. I then wrote a Python script (see Appendix A) using Regular Expressions (RegEx) to take all instances of particular words or word-patterns found in the partial-corpus and replace them with selected gender-neutral words *en-masse*. This would then be placed in a separate text file for training use.

The words I elected to manipulate are listed below, along with the RegEx used to identify them, as well as any interesting notes about their immediate effects on the corpus. In general, the strategy here was to take instances of gendered pronouns or nouns (such as '*he*' or '*woman*') and replace them with gender-neutral ones ('*they*'). Important to note for many of these is that the gender neutral terms often already appear in the original corpus, particularly in reference to groups of people, but in the replacement functions I have made use of the controversial singular '*they*'. To accomodate and easily analyze this, is why I opted to use the version of the corpus tagged with part-of-speech.

For the purposes of understanding the tags used in this list, refer to this list. The RegEx listed here might appear slightly different to the ones in the code (Appendix A), but this is due to Python's string formatting requiring an additional backslash ("\") to escape an actual backslash.

Word Selections:

Conversion	**Identifying RegEx**	**Replacement Text**
He/She -> They	\b([Hh]e|[Ss]he)_PPS\b	they_PPS
He's/She's (is) -> They're (are)	\b([Hh]e's|[Ss]he's)_PPS\+BEZ\b	they're_PPS+BER
He's/She's (has) -> They've (have)	\b([Hh]e's|[Ss]he's)_PPS\+HVZ\b	they've_PPS+HV
He'd/She'd -> They'd	\b([Hh]e'd|[Ss]he'd)_PPS	they'd_PPS

- In the case of '*They'd*', the ending of this has been left intentionally open as variable between '*had'* and '*would'*, and the lasting tag (e.g. *HVD*) will be preserved in-place.

Conversion	**Identifying RegEx**	**Replacement Text**
He'll/She'll -> They'll	\b([Hh]e'll|[Ss]he'll)_PPS\+MD\b	they'll_PPS+MD
Him/Her -> Them	\b([Hh]im|[Hh]er)_PPO\b	them_PPO
His/Hers (Nominal) -> Theirs	\b([Hh]is|[Hh]ers)_PP\$\$\B	theirs_PP$$

- In order of operations, I have intentionally placed the nominal possessive (*PP$$*) ahead of personal possessive (*PP$*) to avoid a conflict where '*His_PP$$'* could be caught and overwritten by identifying '*His_PP$'* inside of it.

His/Her (Personal) -> Their	\b([Hh]is|[Hh]er)_PP\$\B	their_PP$
Himself/Herself -> Themself	\b([Hh]imself|[Hh]erself)_PPL\b	themself_PPL
Man/Woman -> Person	([Ww]oman|(?<![Hh]u)[Mm]an)(?=('s)?_(?=NN))	person

- A specific case has been made here to exclude '*Human_NN*'.
- This replacement and a few others remaining have allowances made for plural nouns (as in *Men's*), such that the part-of-speech tag will be held in-place - so long as the matching phrase is a noun (*NN* will still become *NN*, and *NN$* will be *NN$*).
- This replacement and a few others remaining have allowances made for compound words that involve gendered nouns (e.g. '*Congresswoman'* will become '*Congressperson'*), but these must be non-proper nouns (e.g. avoiding '*Rothman_NP'* or '*German_JJ'*) and must be at the end of the word (e.g. avoiding '*Manufacture'*).

Men/Women -> People	([Ww]omen|[Mm]en)(?=('s)?_(NNS|NPS))	people
Girl/Boy -> Child	([Gg]irl|[Bb]oy)(?=('s)?_(NN))	child
Girls/Boys -> Children	([Gg]irls|[Bb]oys)(?=_(NNS))	children
Girls'/Boys' -> Children's	([Gg]irls'|[Bb]oys')(?=_(NNS\$))	children's

- In the case of these noun swaps, there are a couple of strange instances that are hard to avoid, such as *cowboy* and *Playboy*, but these didn't seem to occur enough to warrant a special case, the way '*human'* does.
- Other instances where these often go wrong are for multi-word idioms, like "*men and women*" feature together often, and now clumsily appear as "*people and people*".
- In the case of compound nouns, these expressions depend on the gendered term being at the end of the word, which means missing out on words like '*Boyfriend*'.

This set of words swaps is of course far from exhaustive, as with more time it would be interesting to pursue a more specific list of compound words (Correcting '*Congresspeople'* to '*Members of Congress'* could be less clumsy, but also very specific) as well as a more extensive list of implicitly gendered nouns (like *'lady'* or '*king'*) or even titles ('*Mrs.'*).

It's also possibly interesting to look into the prospects of all kinds of androcentric terminology (e.g. '*man-made'*), and likewise gynocentric ('*midwife'*) in a broader social evaluation, but in the context of a transformer model it seems unlikely to find relationships based on words that simply have the word *man* in them; we care about words that group with other words in a sentence.

Because of this is why I primarily chose to go with pronouns - it seems that in a sentence describing people and their actions, pronouns often appear to do a lot of the "heavy-lifting" of propagating gender throughout a normal sentence, without generally adding too much to the meaning of the sentence beyond gender; "*She went to her car*" and "*They went to their car*" are factually identical as sentences beyond the absence of gender.

Below is a table outlining the total number of changes actually made to the training corpus (based on running the RegEx search on each term), with the terms put into a shorthand 'pseudocode'. I haven't included the counts for gendered terms in the manipulated corpus, as they have all been removed.

Original Brown Corpus (First 50%)						**Manipulated Corpus**	
Masculine Term	**#**	**Feminine Term**	**#**	**Neutral Term**	**#**	**Neutral Term**	**#**
he_PPS	3908	she_PPS	773	they_PPS	0	they_PPS	4681
he's_PPS+BEZ	33	she's_PPS+BEZ	7	they're_PPS+BER	0	they're_PPS+BER	40
he's_PPS+HVZ	7	she's_PPS+HVZ	1	they've_PPS+HV	0	they've_PPS+HV	8
he'd_PPS+?	9	she'd_PPS+?	1	they'd_PPS+?	0	they'd_PPS+MD	10
he'll_PPS+MD	11	she'll_PPS+MD	2	they'll_PPS+MD	0	they'll_PPS+MD	13
him_PPO	931	her_PPO	249	them_PPO	907	them_PPO	2087
his_PP$$	11	hers_PP$$	3	theirs_PP$$	12	theirs_PP$$	26
his_PP$	3374	her_PP$	631	their_PP$	1610	their_PP$	5604
himself_PPL	293	herself_PPL	29	themself_PPL	0	themself_PPL	322
?man('s)_NN?	916	?woman('s)_NN?	78	?person('s)_NN?	110	?person('s)_NN?	1104
?men('s)_N?	645	?women('s)_N?	120	?people_('s)N?	562	?people('s)_N?	1327
?boy('s)_NN?	122	?girl('s)_NN?	81	?child('s)_NN?	112	?child('s)_NN?	315
?boys_NNS	82	?girls_NNS	65	?children_NNS	198	?children_NNS	345
?boys'_NNS$	3	?girls'_NNS$	1	?children's_NNS$	17	?chlidren's_NNS$	21
TOTAL	**10345**	**TOTAL**	**2041**	**TOTAL**	**3528**	**TOTAL**	**15903**

Most interesting about these numbers is the overwhelming amount of masculine terms in the original corpus, in raw numbers they outnumber feminine terms 5:1. Possibly more interesting still is that even neutral terms actually outnumber feminine terms as well, even after accounting for the fact that neutral pronouns are never used in singular - people are more likely to be assigned as groups than as an individual woman.

Section 2: Comparing Output

Once the corpus was prepared in the previous section, I set about using Max Woolf's Collaboratory Notebook to train two separate models of GPT2-Simple, one with the original partial corpus - 'Baseline', and of course the other with the manipulated corpus, 'Neutralized'. I trained each of these on the '124M' model, for 1000 iterations each. The Baseline model on it's last run achieved a loss value of 1.34, and an average of 1.56, whereas the Neutralized model reached 1.44 loss and 1.56 average.

With each of these trained, I selected a total of 8 gender-coded 'root' words that I would use to generate and analyze the output of. These 8 words are divided evenly into two categories: Gender-Coded Adjectives, and Gender-Coded Occupation Nouns. Within each of these categories are two words for masculine and feminine each. All of these words were taken and placed into some starting sentence structures, and used as a generation seed/prefix, for which each generated five sample texts of 250 words per text.

Below these words have been listed, along with the full seed string used to generate, along with notes on the qualitative features of the generated texts of each. I've tried to select feature words from each sample that stood out as themes, divided into Primary (gender/neutral noun or pronoun words that we are looking for) and Secondary (Any words that stood out as potentially coded along with the theme).

Gender-Coded Adjectives:

For these adjectives I inserted the word in the partial sentence "*Such a ____ person*" to try and prime the attention mechanisms toward describing people (without doing this, words often drifted to objects, e.g. '*Beautiful'* often talked about '*art'*), importantly making sure that the adjective is on describing the person, trying to hint at the position encoding that the transformer should take advantage of. For these I have excluded '*person_NN'* from the primary features, as it's already primed in the prefix.

Beautiful_JJ (Such a beautiful person) - occurs 77 times in each corpus.

- **Baseline:** Four of the five samples developed into describing women almost immediately, the fifth sample ended up being mostly nonsense. Most of these ended up being about their bodies, and sometimes their voices. One sample ended up describing a woman killing a man with her heart. Some samples had a noticeably 'poetic' tone.
    - **Primary Features:** *her_PP$, she_PPS, Mrs._NP, Miss_NP, his_PP$, him_PPO, women_NNS, females_NNS, children's_NNS$*
    - **Secondary Features:** *eyes_NNS, body_NN, strongest_JJT, voice_NN, natural_JJ, heart_NN, sexual_JJ, fancy_NN, attractive_JJ, lovely_JJ, skin_NN, innocence_NN, skintight_JJ*
    - **Verdict:** Overwhelmingly feminine.

- **Neutralized:** The first neutralization seems to have gone very well, zero references at all to any gender directly. Some of the themes have maintained from the baseline - predominantly sexual/anatomical features, but it seems to have discarded some feminine-coding of secondary features as well, particularly in regard to additional adjectives like *'fancy_JJ'* or *'attractive_JJ'*.
    - **Primary Features:** *they_PPS, themself_PPL, them_PPO, child_NN, their_PP$, people_NNS, themselves_PPLS*
    - **Secondary Features:** *sexual_JJ, orgasms_NNS, sex_NN, eye_NN, instrument_NN, artist_NN, naked_JJ, clothing_NN, nose_NN, eyes_NNS, love_NN, slave_NN, cruel_JJ, country_NN, humble_JJ*
    - **Verdict:** Mostly neutral characters, possible feminine secondary coding.

Submissive_JJ (Such a submissive person) - occurs 4 times in each corpus.

- **Baseline:** All of these samples included references to male characters, while four of them included female characters, but this difference in number doesn't paint a full picture, as most of the stories told here feature a man juxtaposed against the *'submissive'* character - more usually the female.
    - **Primary Features:** *women_NNS, he_PPS, his_PP$, wife_NN, child_NN, her_PPO, man_NN, her_PP$, woman's_NN$, woman_NN, hate_VB*
    - **Secondary Features:** *aggressors_NNS, whore_NN, sexual_JJ, sex_NN, marry_VB, married_VBN, loves_VBZ, religious_JJ, divorce_NN, autonomy_NN, bed_NN, attractive_JJ, lover_NN, conquest_NN, affairs_NNS, transvestite_NN*
    - **Verdict:** Balanced, skewing feminine, particularly on characters.

- **Neutralized:** Some gendered themes are still left over in these samples, two of the samples were sexual in nature, which were coincidentally the two that contained remnant gendered language, one including '*man's_NN$'* - interesting because it doesn't even exist in the neutral corpus. The features of men again were juxtaposed against the submissive character though. Remaining themes are more neutral, and take on a more sinister tone with themes about smallness, children, and victimization.
    - **Primary Features:** *they_PPS, man's_NN$, person's_NN$, child_NN, female_NN, father_NN, brother_NN*
    - **Secondary Features:** *sex_NN, breathing_VBG, little_JJ, gentle_JJ, innocent_JJ, victim_NN, victimized_VBN, stereotype_NN, sexual_JJ, hand_NN, arms_NNS, virgin_NN*
    - **Verdict:** Reduced gendering, not eliminated for characters.

Strong_JJ (Such a strong person) - occurs 104 times in each corpus.

- **Baseline:** All of these samples feature male characters, often more than one, and zero female characters or references to the feminine. Overarching themes are more in reference to strength as mental fortitude rather than physical, which is interesting, such as resisting temptation or being of strong conscience.
    - **Primary Features:** *man_NN, his_PP$, him_PPO, he_PPS, himself_PPL, gentleman_NN*

    - **Secondary Features:** *strength_NN, courage_NN, fight_VB, freedom_NN, home_NR, family_NN, conscience_NN, protect_VB, stupid_JJ, temptation_NN, irresistible_JJ, duty_NN, revolver_NN, suspicious_JJ*
    - **Verdict:** Overwhelmingly masculine.

- **Neutralized:** Four of these five samples ended up being about people, most of them making references to "*good_JJ*" people, most of these stories became more varied and descriptive - *'strong_JJ'* has become better able to branch into specific subjects, possibly due to not having such a massive correlation to masculinity that it statistically must refer back to it all of the time.
    - **Primary Features:** *themself_PPL, they_PPSS, they_PPS, children_NNS, child_NN*
    - **Secondary Features:** *football_NN, beliefs_NNS, truth_NN, therapist_NN, hospital_NN, good_JJ, bully_NN, danger_NN, bigger_JJR, firm_NN, gentle_JJ, knight_NN, armor_NN, battle_NN, fighting_VBG, war_NN*
    - **Verdict:** Mostly neutral, somewhat masculine secondary coding.

Dominant_JJ (Such a dominant person) - occurs 18 times in each corpus.

- **Baseline:** Four of the five samples were about male characters (one of which was a dog), no female characters presented. Two stories related very closely to church/congregation. Unlike '*submissive_JJ'* - the opposite, these samples feature no sexual content at all.
    - **Primary Features:** *man's_NN$, his_PP$, they_PPSS, he_PPS, men_NNS, him_PPO*
    - **Secondary Features:** *responsibilities_NNS, leader_NN, attitude_NN, church_NN, expert_NN, freedom_NN, advantage_NN, skill_NN, congregation_NN, education_NN, instinct_NN*
    - **Verdict:** Overwhelmingly masculine.

- **Neutralized:** Most of these samples tended to drift toward themes of power and survival, but interestingly shifted context, two in international geopolitics (centered on the US, as the corpus is), and then two in legal battles. But most importantly, none of them included direct references to gender at all.
    - **Primary Features:** *they_PPS, themself_PPL, they_PPSS*
    - **Secondary Features:** *perfect_JJ, fear_NN, power_NN, influence_NN, foreign_JJ, nation_NN, country_NN, unite_VB, family_NN, trust_VB, debate_NN, legal_JJ, liberty_NN, controls_VBZ, negative_JJ, status_NN, character_NN*
    - **Verdict:** Very neutral, in both characters and secondary coding.

Gender-Coded Occupation Nouns

For these nouns I inserted the the word in the partial sentence "*The* _____ (past-tense verb) *because*", taking a nod from Lu et al. (2020), to nudge to position encoding toward using a pronoun after starting ('*because she…')* to try and make the best demonstration of network's bias of these occupations to one gender or another. Doing this puts the transformer right in the middle of taking a person with an occupation and verb to go off, and probably must assign them a pronoun (and thus, likely a gender) almost immediately, and this actually worked very well in practice. I tried to make sure the verb I picked was at least tangentially relevant to the occupation, and was matched to an equivalent occupation per gender, to control for the verb giving any hints.

Nurse_NN (The nurse cared because) - occurs 8 times in each corpus.

- **Baseline:** All of the samples generated for nurse assigned the nurse immediately as '*she_PPS'*, cementing this as highly feminine-coded almost instantly, but also worth noting is once past the assignment of the nurse themself in these narratives, is that the patients and other characters the nurses interact with actually have a reasonably good split, with a 3:2 split of male:female. Beyond this, the themes presented stayed mostly on topic to medical-care scenarios.
    - **Primary Features:** *she_PPS, they_PPSS, child_NN, him_PPO, he_PPS, boy_NN, her_PP$, man_NN, Mrs._NP, Mr._NP, woman_NN, her_PPO*
    - **Secondary Features:** *procedure_NN, doctors_NNS, plagued_VBN, weight_NN, balanced_VBN, little_JJ, remedy_NN, sensitive_JJ, drug_NN, unconscious_JJ, bed_NN, oral_JJ*
    - **Verdict:** Overwhelmingly feminine (on primary characters at least).

- **Neutralized:** Of these five, all the first instance gender assignments are '*they_PPS*', which is positive, although one of the samples directly assigned the title '*mother_NN'* later on, and another sample didn't explicitly state mother, but did indicate having a baby, possibly indicating maternal secondary themes? That said, another also called the nurse a brave american soldier, so it seems to have balanced out well.
    - **Primary Features:** *they_PPS, person_NN, them_PPO, mothers_NNS, mother_NN*
    - **Secondary Features:** *hospital_NN, evidence_NN, drugs_NNS, treated_VBD, patient_NN, doctor_NN, responsible_JJ, moral_JJ, war_NN, soldier_NN, caretaker_NN, kindest_JJT, lovely_JJ, playmate_NN, nap_NN, baby_NN*
    - **Verdict:** Mostly neutral, possible remaining feminine coding.

Teacher_NN (The teacher studied because) - occurs 37 times in each corpus.

- **Baseline:** Contrary to what I had expected, this set actually had a masculine bias rather than a feminine one, assigning four of the teachers as men, and one as a woman. Interestingly, none of their students had any obvious gender coding.
    - **Primary Features:** *he_PPS, his_PP$, children_NNS, she_PPS*
    - **Secondary Features:** *language_NN, students_NNS, mature_JJ, assignment_NN, instruct_VB, pattern_NN, answers_NNS, classroom_NN, literary_JJ, education_NN, voice_NN, sound_NN*
    - **Verdict:** Heavy masculine bias.

- **Neutralized:** Again, the neutralizer seems to do the job very well, as the very first instance of assigning gender to the teacher is always neutral, and the narratives manage to stick to it. The only mention made of any directly gendered terms are in a single sample, in which the teacher lives with their mother and father - which balances out nicely. Themes stayed on-topic, although with a slightly more religious tint.
    - **Primary Features:** *they_PPS, them_PPO, person_NN, child_NN, their_PP$, father_NN, father's_NN$, mother_NN*
    - **Secondary Features:** *private_JJ, private_JJ, union_NN, schools_NNS, church_NN, student_NN, moral_JJ, compass_NN, God_NP, taught_VBN, School_NN*
    - **Verdict:** Mostly neutral, on both characters and secondary coding.

Doctor_NN (The doctor cared because) - occurs 40 times in each corpus.

- **Baseline:** All five of the samples immediately assigned the doctor as '*he_PPS*', again indicating a masculine bias instantly. That said, once established, the doctor's patients had a nice 50/50 split of male/female. Also interesting in comparison to the nurse samples, the persistent themes present here had much more technical and specific medical language, despite having similar thematic premise and contexts.
    - **Primary Features:** *he_PPS, her_PP$, she_PPS, husband_NN, their_PP$, his_PP$, him_PPO*
    - **Secondary Features:** *patient_NN, bleeding_VBG, neck_NN, diagnosis_NN, radiation_NN, died_VBN, cancer_NN, diabetes_NN, intensive_JJ, hospital_NN, symptoms_NNS, condition_NN, blisters_NNS, fingers_NNS, blood_NN, medical_JJ*
    - **Verdict:** Overwhelmingly masculine (on primary characters at least).

- **Neutralized:** All of the doctor characters here are presented as neutral, although interestingly there is a single sample here where the first word after the prefix is not a pronoun of a person: "*the_AT doctor_NN cared_VBN because_CS it_PPS was_BEDZ the_AT only_AP method_NN available_JJ*", which means the transformer has just this once chosen the verb (*cared_VBN*) as its highest attentional weight when picking its own first word. Also interesting is that in one of the samples, the word '*boy_NN*' appears, another anomaly considering it has been removed from the neutralized corpus entirely.
    - **Primary Features:** *they_PPS, their_PP$, person_NN, child_NN, boy_NN*
    - **Secondary Features:** *patients_NNS, warmth_NN, affection_NN, pills_NNS, die_VB, Hospital_NN, medical_JJ, director_NN, malnourishment_NN, normal_JJ*
    - **Verdict:** Mostly neutral.

Scientist_NN (The scientist studied because) - occurs 14 times in each corpus.

- **Baseline:** All of these samples presented male scientists immediately, and none featured any additional characters beyond the scientist. Aside from this, samples covered a variety of sciences like physics and biology, and stayed on theme once established.
    - **Primary Features:** *he_PPS, man_NN, his_PP$*
    - **Secondary Features:** *discovery_NN, principle_NN, light_NN, heat_NN, universe_NN, atoms_NNS, mass_NN, efficient_JJ, knowledge_NN, time_NN, sound_NN, scientific_JJ, nuclear_JJ, explosion_NN, weapons_NNS, disarmament_NN, biological_JJ, evolution_NN, chemical_NN, theoretical_JJ*
    - **Verdict:** Overwhelmingly masculine.

- **Neutralized:** All scientists here are presented as gender neutral again, and much like the baseline version, contain no other characters than the first scientist. Interestingly, the sciences covered change focus slightly to mainly biology and psychology.
    - **Primary Features:** *they_PPS, they_PPSS*
    - **Secondary Features:** *biologist_NN, paradigmatic_JJ, psychological_JJ, nature_NN, laboratory_NN, scientific_JJ, inspired_VBN, intellectual_JJ, natural_JJ, psychology_NN, physical_JJ, spiritual_JJ, observer_NN, philosopher_NN*
    - **Verdict:** Neutral.

Overall in this, the neutralized model appeared to succeed in its goal about as far as one could reasonably expect, taking almost all instances of gendered pronouns out, so much so that the only remnants of gendered language remaining were from words that simply hadn't been changed, such as *wife_NN* or *son_NN*. The position encodings of pronouns make them very easy and convenient to swap out, as they can be swapped out for the neutral term without changing any of the text meaning - except for the gender of course, which is exactly what was being aimed for anyway.

Interesting also is the subtle effect on secondary themes and features, as when words are embedded in a model as strongly as, say, '*beautiful'* and '*she'*, a transformer building text containing '*beautiful'* will naturally respond with '*she'* rather than '*he'* most of the time if positionally appropriate - assuming no other factors. But this also has a knock-on effect when progressing further into text, as if the same model then progresses onward to generate an occupation from there, and has to build upon '*beautiful'* and '*she'*, it's again more likely to respond with something like '*nurse'* rather than '*doctor'*.

So by replacing all instances of gendered words, we effectively are 'merging' the word-selection spaces of masculine and feminine words in our model, as when primed on '*beautiful*', the model now will only reply with *they* in an appropriate position, and the knock-on effect is that from there all words chosen from that point onward no longer continuously reinforce each other toward femininity, or at least are much less likely to, and must rely on weaker secondary features to do so.

Section 3: As models of 'actual' English

After examining the changes of bias in the models, the next task is to compare the models on how well they perform as models of 'actual' English. As suggested, I tried taking the two trained models, and training them on a new dataset - the second half of the Brown corpus that was held-out initially. I trained each model on this dataset for 10 iterations, to try account for natural variance in training, and averaged the losses and averages for each. The Baseline model achieved an average loss of 1.295, and average average of 1.307, whereas the Neutralizer model had an average loss of 1.311, and an average average of 1.335. This places the Baseline model comfortably as the 'better' model, at least as a model that iteratively reproduces text from the Brown corpus. This is to be expected, fundamentally changing the corpus, particularly by reducing the descriptive power of words as pervasive as gendered-pronouns, makes a model inherently worse as a predictor of text that is more similar to the original than the manipulated version.

This also translates to being a worse model of 'actual' English as well, for a couple of reasons, not the least of which is that for the purposes of this assignment, the Brown corpus is the best approximation of English we have. More than that, the manipulated model is largely incomplete; it can't handle idioms as simple as *'ladies and gentlemen*' without making them sound exceptionally clumsy; '*ladies and gentlepeople*'. When converting contracted pronouns like *'he's'* to *'they're'*, the manipulated model manages seemingly well, but uncontracted '*he is*' converts to *'they is*'.

But probably the biggest reason is that the greatest strength of manipulation like this is also its greatest weakness - removal of the ability to tell genders from one another. Whilst there are plenty of examples of ways in which bias can be harmful when applied to gender constructs, there is a utility we gain from being able to describe aspects of gender that genuinely *are* unique from one another, and losing that also risks another type of bias, akin to when one opts to ignore race and becomes "colourblind" (a la "All Lives Matter"). This loss of descriptive power, whether for better or worse, makes the neutralized model a less accurate predictor of English text, via the Brown corpus or otherwise.

Appendix A: Corpus Manipulation Code (Python)

```
import regex

readfile = open("brown-tagged-part1.txt")
writefile = open("brown-neutralized-part1.txt", "w+")

text = readfile.read()

text = regex.sub("\\b([Hh]e|[Ss]he)_PPS\\b", "they_PPS", text)
text = regex.sub("\\b([Hh]e's|[Ss]he's)_PPS\\+BEZ\\b",
"they're_PPS+BER", text)
text = regex.sub("\\b([Hh]e's|[Ss]he's)_PPS\\+HVZ\\b",
"they've_PPS+HV", text)
text = regex.sub("\\b([Hh]e'd|[Ss]he'd)_PPS", "they'd_PPS", text)
text = regex.sub("\\b([Hh]e'll|[Ss]he'll)_PPS\\+MD\\b",
"they'll_PPS+MD", text)

text = regex.sub("\\b([Hh]im|[Hh]er)_PPO\\b", "them_PPO", text)
text = regex.sub("\\b([Hh]is|[Hh]ers)_PP\\$\\$\\B", "theirs_PP$$",
text)
text = regex.sub("\\b([Hh]is|[Hh]er)_PP\\$\\B", "their_PP$", text)
text = regex.sub("\\b([Hh]imself|[Hh]erself)_PPL\\b", "themself_PPL",
text)

text = regex.sub("([Ww]oman|(?<![Hh]u)[Mm]an)(?=('s)?_(?=NN))",
"person", text)
text = regex.sub("([Ww]omen|[Mm]en)(?=('s)?_(NNS|NPS))", "people",
text)

text = regex.sub("([Gg]irl|[Bb]oy)(?=('s)?_(NN))", "child", text)
text = regex.sub("([Gg]irls|[Bb]oys)(?=_(NNS))", "children", text)
text = regex.sub("([Gg]irls'|[Bb]oys')(?=_(NNS\\$))", "children's",
text)

writefile.write(text)
```