Sally the Congressperson: Interlocutor Agency Modulates the Role of Subjective Social Beliefs in Linguistic Processes

Evidence from English Gender-Neutral Role Nouns

Brandon Papineau
Linguistics, Stanford University
Qualifying Paper 1
in partial fulfillment of the requirements for the PhD in Linguistics

Abstract

There is a growing consensus in the psycholinguistic literature that interlocutors rapidly and predictively integrate both intra-sentential and extra-sentential knowledge into their processing of language. This has been formalized as the notion of 'surprisal' (Hale, 2001; Levy, 2008), whereby a word's processing difficulty is proportional to the negative log probability of that word occurring in a particular sentence frame in a particular context. What has been left relatively understudied, however, is the mechanism by which *subjective* rather than *objective* world knowledge is incorporated into linguistic processes, if such knowledge is incorporated at all. We propose here two possible mechanisms by which this might occur. In the first, this information arrives implicitly via intra-sentential surprisal, the biases themselves being instantiated in the language we use to calculate surprisal. In the second, ideologies about subjective social phenomena are incorporated at some level above surprisal. We examine the processing and production of gender-neutral English role nouns (e.g. *congressperson*) in a pair of web-based experiments. We argue that our results speak to a bipartite way in which subjective social beliefs enter the linguistic system: in processing, these beliefs enter via the biased language to which we are exposed, increasing surprisal on 'unbiased' terms. In production, where language users maintain more agency, interlocutors are additionally able to integrate their subjective social beliefs at a level above what would be predicted by their exposure to the language's biases. We conclude with a brief discussion of the implications of this work, and a call to incorporate more gender-neutral language in the domain of role nouns.

Keywords: language and gender; language processing; language production; language and politics; morphology

Introduction

There is a growing consensus among psycholinguists that individuals rapidly integrate knowledge of the real world in their processing of linguistic input, alongside the linguistic input itself. In their seminal paper, Tanenhaus et al. (1995) demonstrated that participants were able to use visual cues to resolve attachment ambiguities in real-time, as evidenced by a decreased proportion of incorrect looks to a distractor item in a visually-disambiguating context. In a similar vein, Sumner et al. (2014) argues for a dual-route approach to the processing of linguistic and social information in speech, which necessarily includes an interactive and integrated relationship between the two kinds of information. As summarized in McRae and Matsuki (2009), this kind of integration of real-world knowledge can be activated by single words (Ferretti et al., 2001; Hare et al., 2009; McRae et al., 2005) or by 'combined concepts', or phrases that activate multiple event participants who in some way evoke specific knowledge (for example, *wash* can refer to either hair or a car, but *wash* when co-occurring with *hose* more readily activates the car-washing event than the hair-washing event, as in Matsuki et al., 2011).

Additionally, recent evidence has pointed strongly towards the idea that participants are able to integrate this kind of information *predictively*, such that they can use it to anticipate as-yet unreceived linguistic input. Altmann and Kamide (1999) found that participants were able to use semantic selectional restrictions to anticipate the upcoming argument of specific verbs in a visual-world paradigm. Other semantic knowledge which has been found to integrate into predictive sentence processing includes animacy (Warren & McConnell, 2007), event participation roles (Ferretti et al., 2001), and gender (Duffy & Keir, 2004; Foertsch & Gernsbacher, 1997; von der Malsburg et al., 2020), among others.

The predictability of a given word in context has been formalized as surprisal (Hale, 2001; Levy, 2008), under which a word's processing difficulty should be proportional to its surprisal given previous input $w_1, \ldots, w_{i-1}$ and any extralinguistic or extrasentential context C.

$$\text{processing difficulty} \propto -\log P(w_i | w_1, \ldots, w_{i-1}, C) \qquad (1)$$

This account has received ample empirical support: for instance, more contextually surprising words incur greater reading times (Aurnhammer & Frank, 2019; Goodkind & Bicknell, 2018; Monsalve et al., 2012; Smith & Levy, 2013) and more negative N400 amplitudes (Delogu et al., 2017; Frank et al., 2013).

Under this formulation, there are at least two ways in which extra-linguistic information influences processing. In the first, real-world tendencies simply influence the things we discuss and the ways we discuss them. For example, the rules of physics tell us that things usually fall *down*, not *up*. This fact informs the relative infrequency with which the verb 'fall' takes 'up' as a preposition relative to 'down'; real-world knowledge thus drives language use. This in turn means that interlocutors are more likely to be surprised by the relatively infrequent 'up' occurring after 'fall' than they would be by the occurrence of 'down', based on previous exposure to, and expectations of, the language.

If, however, those same interlocutors are playing a video game with a fantasy gravitational system unlike our own, it may become possible to fall *up*. In this (extra-sentential) context, the relative surprisal of 'fall up' is predicted to be lower than in a standard gravitational context. Indeed, experimental evidence indicates that context manipulation modulates intra-sentential surprisal effects for identical utterances (Berkum et al., 1999; Cook & Myers, 2004; Creer et al., 2018). This is then the second route via which real-world knowledge enters the linguistic system in this formulation.

Where this becomes more complicated, however, is when 'world knowledge' concerns contested subjective social ideologies, rather than more 'standard' knowledge. We can exemplify this by extending the previous example. Suppose English speakers are more or less uniform in their knowledge that things fall *down*, not *up*. Because of this, the relative surprisal of these prepositions are likely to be generally stable among the English-speaking population. Compare this with the idea of gender, which is socially constructed and may vary in its conception from individual to individual. The same speakers who agree that things fall *down* may be more divided on whether or not they believe women should fill traditionally masculine-dominated occupations, such as electricians or hunters. Bearing this in mind, the question becomes how individuals' subjective beliefs about social phenomena are integrated into linguistic processes, since the linguistic experience and world-based knowledge related to these phenomena may not correlate the way we expect them to in cases like those of gravity discussed above.

If social ideology enters the linguistic system intra-sententially, or via word co-occurrence probabilities mediated by real-world tendencies, we do not necessarily expect that there will be a processing difference between ideologically-opposed participants on items relevant to these ideologies.[1] On the other hand, if social ideology comes to bear at some level above surprisal, then we expect ideological opponents to perform differently when it comes to processing socially-charged terms.

Following a tradition of investigating processing and production via the lens of gender, as propogated by Duffy and Keir (2004), Foertsch and Gernsbacher (1997), Pozniak and Burnett (2021), and von der Malsburg et al. (2020) and others, we investigate the question of how social ideology enters linguistic processes via this same lens. Departing from these studies, however, our linguistic items of interest were English 'role nouns', which describe individuals' social and professional positions in the world (Misersky et al., 2014). The subset of forms we employed are morphologically marked for gender, allowing us to look specifically at gender-neutral forms such as 'congressperson'.

This focus reflects ideological associations between such forms and gender-progressivism that have been espoused in public discourse. For example, former Acting Director of National Intelligence Richard Grenell tweeted an image of a cookie with an accompanying display-case card that read "Gingerbread Person". Alongside this was Grenell's caption: 'Stop voting for Democrats.' (Grenell, 2021). Grenell explicitly draws on language ideology to implicitly assert that elected Democrats are responsible for the proliferation of politically-correct language regarding gender. As such, these compound forms offer a fertile ground for investigating gender ideologies in linguistic processes.

We investigated how individually-held ideologies affect the processing and production of gender-neutral language in two web-based experiments. The critical forms included both compound forms (n=14) which make a ternary distinction between male, female, and gender-neutral forms, as well as affixed forms which make only a binary distinction (n=6); see (1a) and (1b) for examples.

(1) **Critical Items**

   a. **Ternary:** *congressman, congresswoman, congressperson*

   b. **Binary:** *villain, villainess*

Experiment 1 examined whether, and in what ways, the *processing* of gender-neutral nouns is modulated by individuals' gender ideology. Our results indicate that individuals' ideologies do not significantly impact the processing of gender-neutral terms, lending support to the hypothesis that ideologies are reflected in the language itself and influence processing via exposure to linguistic instantiations of these ideologies.

Experiment 2 examined how gender ideology affects the *production* of these terms, in order to investigate whether the introduction of more interlocutor agency would demonstrate an ideological difference; such a finding would indicate that individual ideologies *do* enter the linguistic system at some level above linguistic exposure, though only when the interlocutor can introduce it agentively. Our results indicate that this is indeed the case; we found that politically left-leaning individuals, who scored higher on gender-progressivism than their conservative counterparts (see Fig. 7), produced more gender-neutral forms. This suggests that subjective ideologies are incorporated in the linguistic system, at least in the domain of production.

We begin with a description of our norming study and materials before turning to the two primary experiments. We finally conclude with a discussion of how these findings contribute to our understanding of the relationship between social ideology and linguistic processes.[2]

[1] One might reasonably argue that our ideologies are reflected in and established by the language we are exposed to, and R. A. Davies et al. (2017) and Yap et al. (2012) have found evidence supporting the idea that individual differences in exposure to linguistic items modulates processing. However, as explicated in the Methodology of Experiment 1, we found relatively similar rates of use of the critical items (described below) across a range of media, indicating that this line of reasoning may not pertain to the present study.

[2] It is important to note that many of the assumptions in our designs, such as the decision to use 'male' and 'female' names, implicitly endorse or perpetuate the notion of gender as a binary. I would like to highlight that these decisions in no way reflect the beliefs or values of the author.

Experiment 0: Norming Study

In order to select role nouns which represented a range of expectations about the gender of their referents, we normed 39 such role nouns for gender probabilities.

Methods

Participants 100 participants were recruited through the online recruitment platform "Prolific" (2014). All participants self-identified as L1 English speakers and as having been born and residing in the United States, and all lived in the United States at the time of participation.

All participants, regardless of final inclusion, were paid $2.00 for their participation in the study, and spent an average of four to five minutes on the experiment (average payout $31.86 per hour).

Stimuli & Procedure The stimuli in this norming study consisted of sentence frames with the structure "Someone is a [ROLE NOUN]". One of 39 role nouns appeared in each of these frames, in one of its possible gendered permutations; 11 of these items were of the binary form in (1b), and the remaining 28 were of the ternary distinction presented in (1a). The full list of these items and their forms can be found in Appendix A. Role nouns were selected from a variety of online guides on avoiding gendered language in English.

For each of the sentences, participants were asked to indicate on a seven-point Likert scale how likely they thought it was that the 'someone' in question was male or female. The Likert scales were randomized between participants with regard to which end of the scale represented a (fe)male response. For each item, participants had the option of selecting a box that said 'I am not familiar with this term'. No participants in the sample opted to select this, indicating the recognizability, if not necessarily high frequency, of the critical items.

No participant saw the same lexeme multiple times with different gender morphology. All participants saw 13 items of each gender (male/female/neutral).

Finally, after completing the Likert task, participants filled out an optional post-experimental demographic survey, including questions about their own gender, political affiliations, and age.

Results

Exclusions Because we expect the female-marked forms to be rated near to ceiling as female-referring, participants were excluded from analysis if their mean score for their female items were less than two standard deviations away from the group mean. This resulted in an exclusion of 11 participants, for a total datset of 89 participants and 3,471 observations.

Gender Ratings The mean gender rating for each of the 106 critical items is presented in Fig. 1; a score of 7 indicates maximal expectation of a female referent, while a score of 1 indicates a maximal expectation of a male referent.

As predicted, the female-marked forms were judged to be the most likely to refer to female individuals, with all items performing at or near ceiling. Most of the neutral forms hovered around 4, indicating equal likelihoods for male and female referents. Notable exceptions included those items with strong male associations (e.g., *firefighter*, *garbage collector*, *fisher*), and items which make only a male-female distinction and whose female forms are relatively frequent (e.g. *emperor/empress*, *actor/actress*).

Finally, of all the neutral forms only one had a score above 5 (indicating a tendency towards female referents): *flight attendant*, a role traditionally and dominantly filled by women.

Discussion

Based on these results, we selected 20 items which ranged from very masculine-coded (*hunter*) to relatively female-coded (*flight attendant*). We excluded items that might be contentious for other reasons (*god*) and those which were uncommon in American English or relatively dated (*head-teacher*, *paperboy*, *door attendant*). The selected items[3] were then used as the critical items for the subsequent experiments.

Experiment 1: Self-Paced Reading

In an experiment similar to that of the processing experiment in von der Malsburg et al. (2020), our first investigation concerned the role that individuals' ideologies about gender play in their processing of gender-neutral role nouns. If participants do exhibit effects of gender ideology, we expect that gender-progressive participants will show faster reading times on gender-neutral terms. If ideology itself does not modulate processing, we expect there to be no difference between individuals' reading times as a function of ideology.

Methods

Participants 298 participants were recruited through the online recruitment platform "Prolific" (2014), excluding any participants who failed to correctly respond to at least 85% of attention check questions (n=19).[4] All participants additionally self-identified as L1 English speakers and as having been born in and currently residing in the United States. Participants were coded to one of three political macro-categories depending on their self-identification of their political beliefs in the post-experimental questionnaire; participants who identified left of center were classed as 'Democrats', while those right of center were classed as 'Republicans'. Those who placed themselves in the middle were classed as 'Non-Partisans'. See Table 1 for participant demographics.

Stimuli & Procedure In a web-based self-paced reading task, participants saw a series of 20 sentence sets of the form "[NAME] is a(n) [TITLE] from [STATE]. S/he likes [ACTIVITY]", where "[TITLE]" stands in for the critical item of gendered role noun. The states and activities were randomized at the stimuli creation stage so that they remained

[3] These items are indicated in Appendix A with an *.

[4] 200 participants were initially recruited, and an additional 98 Republican participants were subsequently recruited after the original sample revealed a heavy skew towards Democrat-identifying participants.

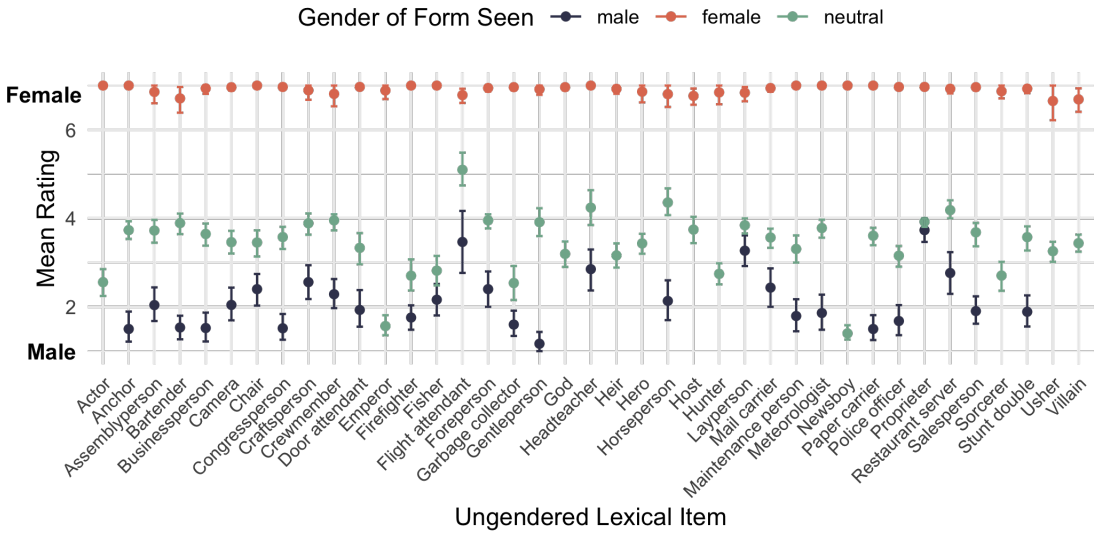

Figure 1: Mean gender value results of the norming study. A score of 7 indicates maximal feminine ratings, whilst a score of 1 indicates a maximal male rating.

Table 1: Experiment 1 and 2 Participant Demographics (Democrat/Republican/Non-Partisan)

	Experiment 1	Experiment 2
Female	64/41/34	82/62/25
Male	46/59/25	42/46/10
Other	3/0/0	4/0/0
Decline to state	0/3/1	1/0/1

constant for all participants. Names varied such that each participant saw 10 vignettes with male-coded names and 10 with female-coded names. Role nouns were then distributed so that 5 of the female names co-occurred with female-marked forms and the other 5 with neutral forms; the same was true for the male names, but with male-marked forms. We intentionally avoided gender-incongruent forms such as 'David is a congresswoman', for fear that doing so would bring too much attention to the research question regarding gender. The resulting conditions are presented in (2) through (5); participants saw each of these combinations five times, followed by activity preferences, for a total of twenty trials. Each name and title occurred only once, such that, for example, no participant saw both *congressman* and *congressperson*.

(2) ***Female congruent***

a. Sally is a congresswoman from Kansas. She likes dancing.

(3) ***Female neutral***

a. Sally is a congressperson from Kansas. She likes dancing.

(4) ***Male congruent***

a. David is a congressman from Kansas. He likes dancing.

(5) ***Male neutral***

a. David is a congressperson from Kansas. He likes dancing.

In order to attain sufficiently-gendered names, the twenty most popular male and female names were selected from the lists of most popular names for boys and girls in 1998 according to the United States Social Security Administration (2021). Names which appeared within the top 100 entries on both lists (e.g. Taylor, Ryan) were excluded.

Participants proceeded through these sentences one word at a time by pressing the spacebar to the reveal the next word and hide the previous; measurements of reading time were taken for each word in the sentence as a proxy for processing difficulty or effort, as has been standardized in the field (Forster et al., 2009). At the end of each trial, participants were asked about properties of the character described, providing a 'yes' or 'no' answer to questions about their home state (*Is Sally from Kansas?*) or about their preferred activities (*Does David enjoy dancing?*); these questions served both to distract from the principal question under investigation, and as attention checks. Participants were provided with an example that did not mark gender before proceeding to the main set of 20 vignettes.

Post-Experimental Survey Upon completing the reading task, participants proceeded to the post-experimental survey.

In order to assess the participants' ideologies towards gender, we employ the Social Roles Questionnaire developed by Baber and Tucker (2006). This survey consists of 13 questions which are designed to elicit both implicit and explicit ideologies about gender, including the notions of gender as an immutable fact vs gender as a social construct (what Baber and Tucker term 'gender transcendence'), as well as about the societal roles performed by the (binary) genders ('gender linking').

Each of the 13 questionnaire items was presented alongside a sliding scale from 'strongly disagree' to 'strongly agree', which corresponded to numerical values of 0 and 100, respectively. The questions related to 'gender linking' were inversely coded and then converted to the same scaling as the 'gender transcendence' subscale. Participants were then assigned a gender ideology score from 0 to 100 by taking the mean of their individual responses; a score of 0 indicated a maximally open-minded approach to gender. The full list of items included in this questionnaire is provided in Appendix B.

Participants also filled out the same demographic questionnaire as in Experiment 0. Participants who declined to indicate their age or political orientation were excluded from analysis.

Unigram Surprisal In order to account for effects of word surprisal, the unigram surprisal of each of the twenty critical items' neutral forms was computed from the 'Spoken' (news media) section of COCA (M. Davies, 2008-). The decision to use unigram, contextless surprisal values was due to the difficulty in obtaining surprisal values for very infrequent terms, such as *foreperson*. The decision to use the same surprisal values for all participants stems from the high correlation between unigram surprisal values in the right- and left-wing sources in COCA (cor = .83).

Results

Exclusions In addition to the aforementioned participant exclusions, 238 trials (4.2%) with response times more than 2.5 standard deviations from that lexical item's mean reading time were excluded.

Sensitivity to Gender Before continuing, it is worth noting that there is both quantitative and qualitative evidence to maintain that participants were both (1) cued to gender by the names they saw, and that they were (2) *not* so sensitive to gender in the task that they were able to ascertain the nature of the question under investigation.

With regards to (1), there is numerical evidence that participants were in some way clued in to the difference between male and female names in the prefix before a particular lexical item. While the effect of referent gender did not come out as a main effect in the model reported below, individual lexical items show a high degree of variability between gendered names. The most striking of these is the lexical item 'flight attendant', which was processed more readily when it appeared after a female name than it was after a male name (see Fig. 2). This is consistent with the results of the norming study, where 'flight attendant' was the only item that showed a distinct skew towards a female interpretation. Such a difference indicates that participants were using gender information in their processing of the critical items.

Insofar as assuring ourselves that participants were not aware of the nature of the task they were completing, we turn to the qualitative domain of experimental comments. Of the 76 participants who provided feedback, only one invoked gender in any meaningful way, espousing their view that women were physically weaker than men, and this meant that men should perform certain occupations that women should not. No other participants invoked gender, except for two who explicitly questioned the relationship between the self-paced reading task and the Social Roles Questionnaire. This, taken in conjunction with the fact that the majority of participants in the production task (Experiment 2) *did* overtly comment on gender in language, we feel certain that the participants in our task were not explicitly attending to gender, and that any effects we see are the result of implicit gender activation.

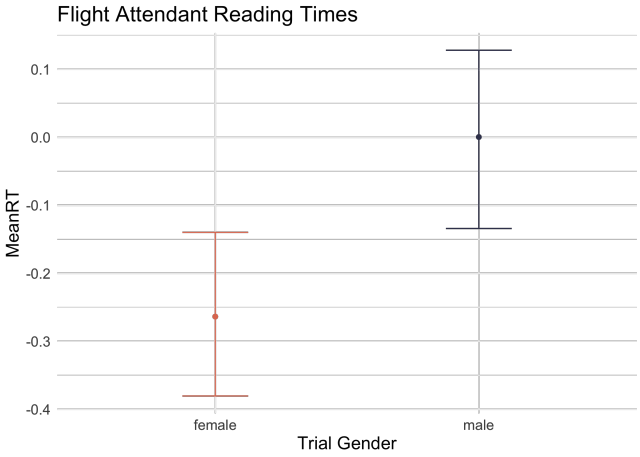

Figure 2: Residualized reading on the item 'flight attendant', by gender of name in the intra-sentential prefix.

Model Structure We fit a linear mixed effect model which predicted length-residualized reading time on neutral terms from fixed effects of political party (ternary, reference level: "Democrat"), referent gender (binary, reference level: "female"), participant age, gender ideology, and unigram surprisal. Random intercepts were included for participant and lexeme. Interactions were included between: ideology and age; surprisal and party; age and surprisal; age and party; ideology and party, and the three-way interaction between age, surprisal, and party. These interactions were included as a result of initial investigations which revealed a significant modulation of surprisal effects by age (Figure 3).

Gender Ideology There was no effect of gender ideology for Democrats (β = -0.00, SE = 0.00, t = -0.11, $p > 0.5$), or in the higher-level interactions for Republicans (β = -0.00, SE = 0.00, t = -1.51, $p > 0.1$) or Non-Partisans (β = -0.00, SE = 0.00, t = -1.34, $p > 0.1$). There was thus no evidence that ideological beliefs about gender and its binary social roles modulate the processing of gender-neutral language. This is similar to the von der Malsburg results, which found no processing advantage for the pronoun which co-referred with the real-world gender of the expected election winner (von der Malsburg et al., 2020).

Political Affiliation At the party-level, we observe no significant difference in reading times on neutral items (location 4 in Fig. 3) between Democrats and Non-Partisans (β = -0.09, SE = 0.4, t = -.26, $p > 0.5$), or between Democrats and Republicans (β = 0.11, SE = 0.33, t = 0.33, $p > 0.5$). These results suggest that party affiliation does not significantly modulate processing of gender-neutral role nouns, either as a result of exposure to different linguistic input or attitudes.

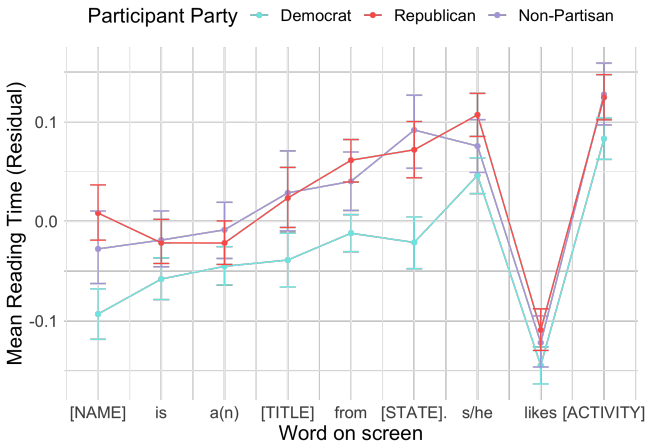

Figure 3: Residualized reading time by word in sentence. "[TITLE]" indicates the location of the critical items.

Unigram Surprisal Mean residualized reading times on neutral terms are shown as a function of political affiliation, age, and surprisal in Fig. 4. More surprising words were read only marginally more slowly overall (β = -0.02, SE = 0.01, t = -1.825, p = 0.07). However, there was a significant two-way interaction between surprisal and participant age, such that older participants showed sensitivity to word surprisal in the expected direction, while young participants did not (β = 0.00, SE = 0.00, t = 2.38, p = 0.018). A 3-way interaction between age, surprisal, and the Non-Partisan party contrast suggests that Non-partisan participants were not sensitive to surprisal (β = -0.00, SE = 0.00, t = -2.54, p = 0.01).

Discussion

The results of Experiment 1 largely support the idea presented in the Introduction that social biases and ideologies enter the

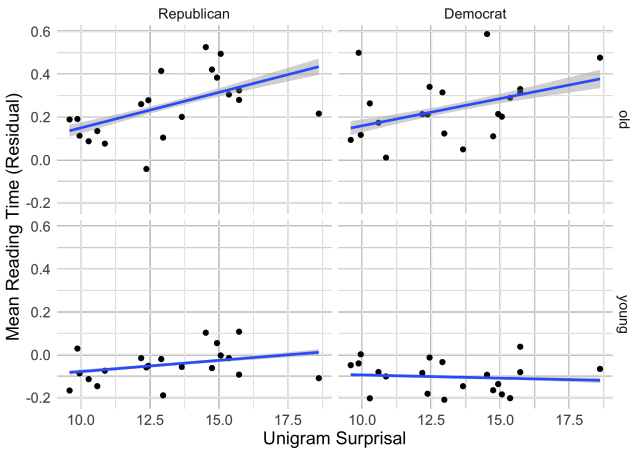

Figure 4: Residualised reading time on critical words by word surprisal, separately for Republicans (left) vs. Democrats (right) and for older (top, >40 years) vs. younger (bottom, ≤40 years) participants. Each point indicates a lexeme.

language via linguistic, not extra-linguistic, pathways. Had there been some affect of gender ideology at a level above the surprisal of the gender-neutral terms occurring in these contexts, we would have expected to see a difference in the processing of these terms between ideologically opposed groups. The results of both gender ideology and political parties indicate that this is not the case. It may be that ideologically-opposed individuals *do* have different surprisal values as a function of the media they consume. However, at least in the domain of English role nouns, this appears not to be the case, considering both that the surprisal effect held for both older Democrats and Republicans, and that the relative frequency of the terms under investigation were similar between the two political wings of the media (Cor = .83).

Moreover, the result that younger participants, regardless of ideology or political party, show little-to-no effect of surprisal compared to their elder counterparts merits further discussion. This finding may indicate that the frequency values obtained from COCA are not representative of the linguistic input experienced by younger Americans. This would be consistent with recent findings that younger Americans are less likely to engage with traditional news sources or trust them, and that they engage with news media in ways not necessarily represented by the data in COCA, such as via Snapchat, Instagram, Reddit, etc. (Antunovic et al., 2018; Edgerly et al., 2018; Kohut, 2013; Madden et al., 2017). Indeed, this claim itself seems to support a theory under which individuals' ideologies themselves do not modulate processing; rather, it is the language to which one is exposed that drives relative processing difficulty in the case of socially-charged language; this is the same argument put forth in von der Malsburg et al. (2020), and our results appear to support it.

However, as also observed in von der Malsburg et al. (2020), there is the possibility that production of gendered

terms will vary even when they do not modulate processing ease. This may be because individual ideologies enter the linguistic system when the language user has more agency and opportunity to ensure that the language they produce concords with their worldviews.

Experiment 2: Forced-Choice Production

As such, we also investigated the role of gender ideology on the *production* of gender-neutral role nouns. In a forced-choice task, participants selected the form of the lexeme they felt best completed the vignettes from Experiment 1. An incongruity in results would on the basis of ideology or political affiliation would indicate that gender ideology is being integrated into the linguistic system in agentive tasks. If ideology is driving lexical choice at some level above the input received by language users, we expect that gender-progressive participants should produce a higher rate of gender-neutral role nouns than their more gender-conservative counterparts.

Methods

Participants 301 participants were recruited using Prolific, with the same criteria as Experiment 1[5]. Participants who failed to correctly respond to 80% of attention checks were excluded (n=25). See Table 1 for participant demographics.

Stimuli & Procedure All items in the experiment consisted of a complete sentence missing a single word, using the same sentence frames and critical items as in Experiment 1. Participants were asked to select the word which best completed the sentence, by choosing from a series of buttons that provided either two or three words. On critical trials, the choice was between the words investigated in Experiment 1.

Filler items took one of two forms; semantic fillers and grammatical fillers. Semantic fillers made use of semantically-related items, all of which resulted in grammatically acceptable sentences, as in (6) and (7).

(6) That's the cutest (horse/Lusitano/equine) I have ever seen!

(7) Revati is a (writer/journalist/author) from India.

Grammatical fillers, on the other hand only one answer which resulted in a grammatically acceptable answer, and employed grammatical processes such as demonstrative selection (8), verb agreement (8), or preposition selection (10), among others. These items served a secondary purpose as attention check questions.

(8) She is typing on (**the**/these/those) computer.

(9) Katherine (**sang**/song/sing) that song beautifully.

(10) They are they eating their soup (between/**with**/at) a spoon.

The presented order of response possibilities was shuffled between participants. There were a total of 80 trials, with 20 critical items and 60 filler items.

Trial order was randomized. After completing the experiment, participants completed the same post-experimental questionnaire as in Experiment 1, and the same demographic questionnaire as in Experiments 0 and 1.

Expectation of Neutrality To control for the possibility that participants simply produce predictable words when faced with a choice, we calculated a neutrality expectation score for each item based on intra-sentential gender reference and word probabilities. Because participants were presented with both gendered and gender-neutral options, we calculated this expectation as the log-transformed relative probability of a neutral noun over a gendered noun occurring, relative to the gender of the sentential subject referent:

$$\text{neutrality} = \log \frac{P(w_{\text{neutral}})}{P(w_{\text{gendered}})} \tag{2}$$

For example, in the sentence 'Sally is a congress[person/woman/man]', the expectation for 'congressperson' is calculated based on the relative probability of 'congressperson' over 'congresswoman'. In contrast, for 'David is a congress[person/woman/man]', the computation is based on the relative probability of 'congressperson' over 'congressman'. These scores make specific predictions: 'police officer' was more frequent than 'police woman' in the corpus, which would predict a neutral response in the female referent gender vignettes. In contrast, 'businessman' was more frequent than 'businessperson', predicting a gendered response on the male referent gender trials.

It should be noted that this calculation assumes prima facie that participants will not choose the 'gender-incongruent' forms 'David is a congresswoman' or 'Sally is a congressman'. As we discuss below, this prediction held true for the male characters. For the female characters, however, it did not. This is discussed in further detail below.

Results

The proportion of neutral and gendered (male, female) noun roles selected are shown in Fig. 5.

Exclusions 241 responses were excluded from analysis for being incongruent with the names that appeared in the vignettes, such as 'David is a congresswoman' or 'Sally is a congressman'. For completeness, these responses are included in Fig. 5.

Model Structure We fit separate logistic mixed effects models for each of the political parties, for the sake of interpretability of interaction terms. These models predicted neutral over gendered responses from fixed effects of neutrality expectation, gender ideology (centered), and referent

[5] 100 Democrats and 100 Republicans were recruited initially, in order to maintain a political balance. An additional 100 male-identifying participants were subsequently recruited due to a significant gender imbalance in the initial participant population (13.4% male-identifying participants in the original population), as a result of an influx of female participants after Prolific went viral on social media app TikTok (Charalambides, 2021).

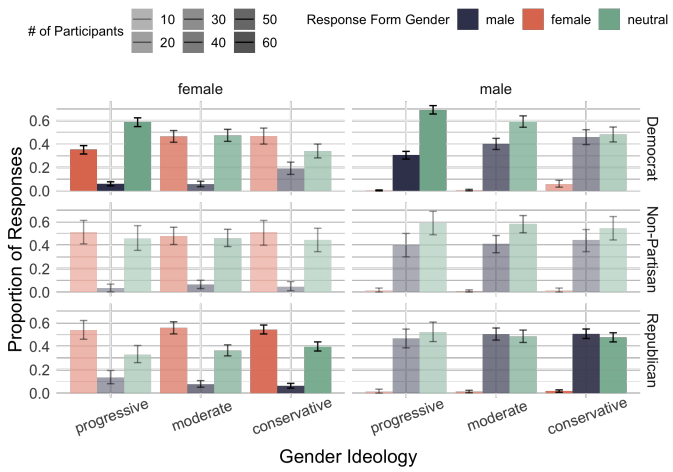

Figure 5: Proportion of neutral and gendered (male, female) responses selected in Experiment 2 as a function of participant gender ideology, separately by referent gender (left: female referents; right: male referents) and participant political affiliation (rows).

gender (binary, centered and scaled, reference level: "male") and the interaction between gender ideology and referent gender. We also included random intercepts for participant and lexical item. The interaction between ideology and referent gender did not reach significance at $p < .05$ for any of the parties. Significant effects are shown in Table 2.

Gender Ideology There was no main effect of gender ideology on the proportion of gender-neutral responses selected (Table 2, Row 2). However, we find that more gender progressive Democrats were more likely to produce gender-neutral role nouns than their less progressive counterparts. Republicans and Non-Partisans showed no such modulation by gender ideology.

Moreover, a mixed effects model on the whole dataset predicting neutral selections from only a fixed effect of political affiliation, with random intercepts for participant and lexical item, showed that Democrats had a higher base production rate of gender-neutral role nouns than their Non-Partisan (β = -0.38, SE = 0.19, z = -2.021, p = 0.04) and Republican (β = -0.83, SE = 0.13, z = -6.4, $p < .001$) counterparts. While Democrats selected the gender-neutral forms 59.6% of the time, Republicans selected them only 45.1% of the time. The Non-Partisans selected the neutral forms at an intermediate rate, 53% of the time.

Referent Gender There was a main effect of sentential referent gender on production rates of gender-neutral titles, such that participants of all three political macrocategories were more likely to produce gender-neutral forms when picking a role title that co-referred with a male name (Table 2, Row 1). Gender-neutral forms were produced 57% of the time with male names, compared to only 48.7% of the time with female names.

Neutrality Expectation Finally, there was an effect of neutrality expectation in the expected direction for all three political parties, such that items with an a priori more frequently used neutral form elicited more neutral responses (Table 2, Row 3).

Discussion

The fact that ideology and political party both significantly modulate the production of gender-neutral forms indicates that ideology is incorporated into the linguistic system at a level above linguistic experience. This is in addition to the fact that gender neutrality expectation, or the calculation of relative probability of a neutral form over a gendered one, also impacts production in the expected direction. The results of our production experiment thus provide evidence that individuals' gender ideologies *do* modulate linguistic processes, when interlocutors have the agency to integrate them and make linguistic decisions that they feel reflect their values and belief systems, even when they are also modulated by relative frequencies.

To this end, it may also be that individuals use these terms as markers of their progressive ideologies, indexing and establishing particularly progressive personae. This, in turn, is likely how associations such as that espoused by Grenell arise. As this is not the main focus of this study, however, we leave this for later investigation and elaboration.

While they were excluded from statistical analysis, it bears mentioning the striking finding that female referents were able to be described by male terms, while male names were almost never assigned female-marked role nouns. This is likely a result of the willingness to treat maleness as the default, the same ideological process by which items like 'villain' and 'actor' have come to be used for both male and female referents, and for referents who identify outside of this binary. This line of inquiry would benefit from a diachronic corpus analysis of these terms, to see if their usages have become less gendered over time.

It is also worth noting that participants were more likely to assign neutral terms to male names than to female names. This may be a case of marked gender-role configurations (i.e. female-coded characters performing roles which were majority male-marked in the norming study) receiving marked descriptors, as the female forms are generally either less frequent (frequency-marked) or morphologically more complex (morphologically marked). To support this claim, we again turn to the female-normed 'flight attendant'. The pattern for this item is crucially the opposite of what we observe elsewhere: 'flight attendant' is more likely in its male condition to be assigned a gender-neutral role noun than with a female coreferent, as in Fig. 6. This being the case, it appears that participants are more likely to choose gender-neutral forms when the gender of the referent does not concord with expectations about the gender of the item.

Table 2: Model outputs for each fixed effect (rows) for each of the political macrocategories.

	Democrats				Non-Partisans				Republicans			
	β	SE	z	p	β	SE	z	p	β	SE	z	p
referent gender	0.86	0.12	6.95	<0.001	1.04	0.22	4.67	<0.001	1.27	0.14	8.92	<0.001
ideology	-0.03	0.01	-4.64	<0.001	-0.01	0.02	-0.37	0.71	0.00	.01	.14	0.89
neutrality	9.23	2.22	4.16	<0.001	15.24	4.62	3.3	<0.001	14.6	2.32	6.3	<0.001

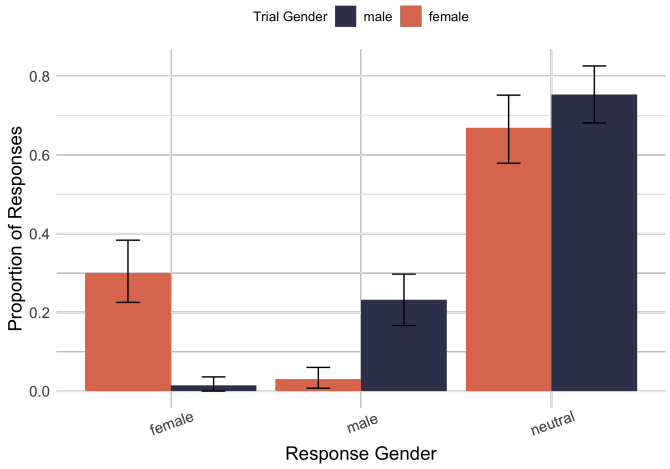

Figure 6: Proportion of neutral and gendered (male, female) responses selected in Experiment 2 on the item 'flight attendant' as a function of the gender of the coreferential name.

General Discussion

We observed no effect of gender ideology on the processing of gender-neutral role titles when they co-referred with gendered names. This is reminiscent of the findings of von der Malsburg et al. (2020), wherein co-referring *she* with *president* incurred a processing penalty despite societal expectations that Hillary Clinton would win the 2016 election. Our data similarly indicates individually-held beliefs about gender do not modulate the processing of gender-neutral role nouns. This seems to suggest that individuals' ideologies do not enter the linguistic system via some system above surprisal in the case of processing. Rather, the implicit biases we hold as a society are reinstantiated in the linguistic system, resulting in processing penalties for surprising genders in particular social roles irrespective of individuals' ideologies.

However, we did observe a difference in processing as a function of age and word surprisal, such that young participants showed less sensitivity to surprisal effects than older participants. This runs counter to previous findings, which have found stronger effects of word predictability in younger participants than in older ones (Moers et al., 2017; Rayner et al., 2006; Steen-Baker et al., 2017). This finding may mean our surprisal values are not accurate for the younger participants in our study, possibly reflecting exposure discrepancies. If this is the case, it would lend further credence to our argument: what influences one's processing of a particular socially-charged item is not their individual ideology, but rather their linguistic experience with that item. This would be better explored and elucidated through additional investigation, possibly making use of media sources known to be the source of linguistic input for younger Americans.

When it comes to the domain of production, however, the results are in striking contrast to those of the processing study. Not only do we find that ideology does modulate production of terms, we find that this is recursively instantiated in the Democratic party, such that more left-leaning Democrats produce a higher proportion of gender-neutral forms than their more conservative counterparts. This modulation seems to indicate that individual ideology is introduced into agentive linguistic processes such as production, or at least selection between semantically-related lexical items.

Taken together, our results thus indicate two avenues by which ideologies about subjective social phenomena enter the linguistic system, though it makes sense to reverse the order they are presented in here. In production, interlocutors maintain a high degree of agency, allowing them to make lexical decisions that reflect their individually-held convictions and ideologies. They are not immune to frequency effects, of course; we find that our participants are still sensitive to frequency effects, as reflected in the main effect of Neutral Expectation in Experiment 2. These frequency effects are themselves the second pathway by which ideology enters the linguistic system and reflect the biases we all experience in our exposure to language. This is evident from Experiment 1, wherein individuals' ideologies did not come to bear on processing, but frequency did. Additional circumstantial evidence comes in the form of age effects, which indicate that linguistic experience facilitates and modulates processing.

This all means that the frequency effects we see are in some way self-fulfilling. The more one encounters a term, the more one may be likely to use it. In turn, that usage of the term adds to the collective knowledge of language users, and users become proliferators of that same form. Dangerously, this also means that these relationships are learned by large language models trained on natural language corpora, raising concerns about the perpetuation of societal biases in the realm of automation and language (Bender et al., 2021; Caliskan et al., 2017; Sutton et al., 2018). Even more alarmingly, in the case of language models' productions, our models are not capable

of integrating ideologies to mediate some of the damaging biases they perpetuate. As such, it becomes evident that the use and proliferation of gender-neutral forms is crucial for the dismantling of gender biases in language. As von der Malsburg et al. (2020) argue, the linguistic system is itself a site of bias perpetuation. In order to overcome these biases, we must employ the agentivity we have in production to eventually normalize these terms in processing.

In sum, we believe that these results further our understanding of the relationship between gender and language by highlighting an incongruity in the processing and production of gender-neutral role nouns, revealing two pathways by which social ideologies enter the linguistic system. Moreover, this incongruity is found at the individual level, calling for a greater degree of granularity in our investigations of biases in the linguistic system, which are critical in the development of fair and inclusive language.

Acknowledgments

This work would not have been possible without the help, support, and insights of more folks than I can name in this space. I would be remiss, however, not to mention the valuable insights provided by my committee: Judith Degen, Robert J Podesva, and Meghan Sumner. I would also like to extend my thanks to Adolfo Hermosillo, Alexia Hernandez, Bonnie Krejci, Brandon Waldon, Chantal Gratton, Christian Brickhouse, Elisa Kreiss, Evelyn Fernández-Lizárraga, Lewis Esposito, Madelaine O'Reilly-Brown, Revati Thatte, Sahba Mobini, and Sarang Jeong. Special thanks to Anthony Velasquez and Stefan Pophristic for innumerable reasons, of which which they are hopefully aware. While all of these folks have contributed to this paper in some way, shape, or form, any remaining mistakes or shortcomings are my own.

References

Administration, S. S. (2021). Popular names in 1998. https://www.ssa.gov/cgi-bin/popularnames.cgi

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Antunovic, D., Parsons, P., & Cooke, T. R. (2018). 'checking'and googling: Stages of news consumption among young adults. *Journalism*, *19*(5), 632–648.

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198.

Baber, K. M., & Tucker, C. J. (2006). The social roles questionnaire: A new approach to measuring attitudes toward gender. *Sex Roles*, *54*(7-8), 459–467.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Berkum, J. J. v., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the n400. *Journal of cognitive neuroscience*, *11*(6), 657–671.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Charalambides, N. (2021). We recently went viral on tiktok - here's what we learned. https://blog.prolific.co/we-recently-went-%20viral-on-tiktok-heres-what-we-learned/

Cook, A. E., & Myers, J. L. (2004). Processing discourse roles in scripted narratives: The influences of context and world knowledge. *Journal of Memory and Language*, *50*(3), 268–288.

Creer, S. D., Creer, S. D., Cook, A. E., & O'Brien, E. J. (2018). Competing activation during fantasy text comprehension. *Scientific Studies of Reading*. https://doi.org/10.1080/10888438.2018.1444043

Davies, M. (2008-). The corpus of contemporary american english (coca). https://www.english-corpora.org/coca/

Davies, R. A., Arnell, R., Birchenough, J. M., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1298.

Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and erps. *Cognition*, *161*, 46–59.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559.

Edgerly, S., Vraga, E. K., Bode, L., Thorson, K., & Thorson, E. (2018). New media, new relationship to participation? a closer look at youth news repertoires and political participation. *Journalism & Mass Communication Quarterly*, *95*(1), 192–212.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, *44*(4), 516–547.

Foertsch, J., & Gernsbacher, M. A. (1997). In search of gender neutrality: Is singular they a cognitively efficient substitute for generic he? *Psychological science*, *8*(2), 106–111.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, *41*(1), 163–171.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts n400 amplitude during reading.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, 10–18.

Grenell, R. (2021). Stop voting for democrats. https://twitter.com/richardgrenell/status/1471502835682480128?s=21

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the north American chapter of the association for computational linguistics*.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, *111*(2), 151–167.

Kohut, A. (2013). Pew research surveys of audience habits suggest perilous future for news.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Madden, M., Lenhart, A., & Fontaine, C. (2017). How youth navigate the news landscape.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 913.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & cognition*, *33*(7), 1174–1184.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and linguistics compass*, *3*(6), 1417–1429.

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulikova, A., Öttl, A., et al. (2014). Norms on the gender perception of role nouns in czech, english, french, german, italian, norwegian, and slovak. *Behavior research methods*, *46*(3), 841–871.

Moers, C., Meyer, A., & Janse, E. (2017). Effects of word frequency and transitional probability on word reading durations of younger and older speakers. *Language and Speech*, *60*(2), 289–317.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408.

Pozniak, C., & Burnett, H. (2021). Failures of gricean reasoning and the role of stereotypes in the production of gender marking in french. *Glossa: a journal of general linguistics*, *6*(1).

Prolific. (2014). https://www.prolific.co

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and aging*, *21*(3), 448.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Steen-Baker, A. A., Ng, S., Payne, B. R., Anderson, C. J., Federmeier, K. D., & Stine-Morrow, E. A. (2017). The effects of context on processing words during sentence reading among adults varying in age and literacy skill. *Psychology and aging*, *32*(5), 460.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in psychology*, *4*, 1015.

Sutton, A., Lansdall-Welfare, T., & Cristianini, N. (2018). Biased embeddings from wild data: Measuring, understanding and removing.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

von der Malsburg, T., Poppels, T., & Levy, R. P. (2020). Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 united states and 2017 united kingdom elections. *Psychological science*, *31*(2), 115–128.

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic bulletin & review*, *14*(4), 770–775.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the english lexicon project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53.

Appendix A: Stimuli Materials

Role Nouns

All items included here were used in the norming study (Experiment 0). Items marked with an * were included in Experiments 1 & 2, as well.

- Ternary Distinction Role Nouns
  - *Anchor/anchorwoman/anchorman
  - Assemblyperson/assemblywoman/assemblyman
  - Bartender/barmaid/barman
  - *Businessperson/businesswoman/businessperson
  - *Camera operator/camerawoman/cameraman
  - Chair/chairwoman/chairman
  - Door attendant/doorwoman/doorman
  - *Congressperson/congresswoman/congressman

- *Craftsperson/craftswoman/craftsman
- *Crewmember/crewwoman/crewman
- *Firefighter/firewoman/fireman
- Fisher/fisherwoman/fisherman
- *Flight attendant/stewardess/steward
- *Foreperson/forewoman/foreman
- Garbage collector/garbagewoman/garbageman
- Gentleperson/gentlewoman/gentleman
- Headteacher/headmistress/headmaster
- Horseperson/horsewoman/horseman
- *Layperson/laywoman/layman
- Mail carrier/mailwoman/mailman
- Maintenance person/handywoman/handyman
- *Meteorologist/weatherwoman/weatherman
- Paper carrier/papergirl/paperboy
- *Police officer/policewoman/policeman
- Proprietor/landlady/landlord
- Restaurant server/waitress/waiter
- *Salesperson/saleswoman/salesman
- *Stunt double/stuntwoman/stuntman

- Binary Distinction Role Nouns
  - *Actress/actor
  - Empress/emperor
  - Goddess/god
  - *Heiress/heir
  - *Heroine/hero
  - *Hostess/host
  - *Huntress/hunter
  - Newsgirl/newsboy
  - Sorceress/sorcerer
  - Usherette/usher
  - *Villainness/villain

Gendered Names

These names were used in Experiments 1 & 2.

- Female Names
  - Emily
  - Hannah
  - Samantha
  - Sarah
  - Jessica
  - Madison[6]
  - Elizabeth
  - Alyssa
  - Kayla
  - Megan
  - Lauren

- Male Names
  - Michael
  - Jacob
  - Matthew
  - Joshua
  - Christopher
  - Nicholas
  - Andrew
  - Austin
  - Joseph
  - David
  - William

[6]It is worth noting that the only sitting congressperson with the name 'Madison' at the time of the study was Congress*man* Madison Cawthorn of North Carolina. However, because Madison was only recorded in the top 100 female names from 1998, it was included as a female name in this study.

Appendix B: Social Roles Questionnaire

The following are the items presented to participants in the post-experimental phase of the task, taken from Baber and Tucker (2006). Participants were asked to respond on a sliding scale from 'Strongly Agree' (0) to 'Strongly Disagree' (100). Items (1) – (5), inclusive, were reverse coded.

- **Gender Transcendence**

1. People can be both aggressive and nurturing, regardless of sex.
2. People should be treated the same, regardless of their sex.
3. The freedom that children are given should be determined by their age and maturity level and not by their sex.
4. Tasks around the house should not be assigned by sex.
5. We should stop thinking about whether people are male or female and focus on other characteristics.

- **Gender Linking**

6. A father's major responsibility is to provide financially for his children.
7. Men are more sexual than women.
8. Some types of work are just not appropriate for women.
9. Mothers should make most decisions about how children are brought up.
10. Mothers should work only if necessary.
11. Girls should be protected and watched over more than boys.
12. Only some types of work are appropriate for both men and women.
13. For many important jobs, it is better to choose men instead of women.

Appendix C: Gender Ideology Scores

Fig. 7 presents the ideology scores from all participants across both Experiments 1 & 2. A score of 100 is maximally conservative with regards to gender ideology, while a score of 0 is maximally progressive.

Consistent with public discourse surrounding and conceptions of the two political parties, we find that Republicans reported more conservative gender ideologies than their Democratic counterparts, with self-described moderates performing somewhere in the middle. Republicans additionally show the greatest amount of intra-party variation, while Democrats show the least.

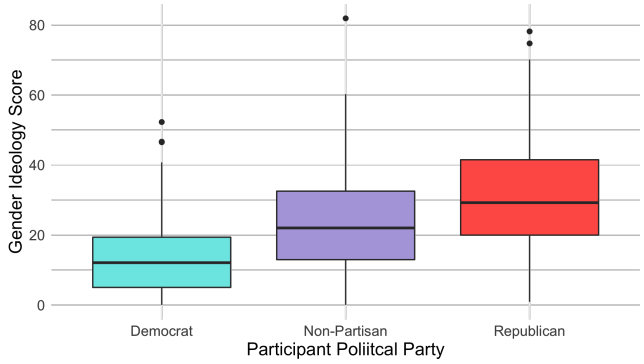

Figure 7: Gender Ideology Scores by Participant Political Party