DOES GENDER BIAS AFFECT THE INTERPRETATION OF ENGLISH SLOPPY ELLIPSIS?*

Dennis Ryan Storoshenko and Jesse Weir
University of Calgary

In this paper, we present the results of a contextual felicity study examining the interpretation of English reflexives *himself* and *herself* under coordinated VP ellipsis. Our findings support existing claims that sloppy readings of the reflexive are the default, while refuting claims that under conditions where an elided *herself* is taken to be bound by a mismatching (masculine) antecedent, a strict reading is facilitated. Adding the gender identity of participants as a variable in the study does not change the main claim regarding the mismatches, but an unexpected interaction emerges where female- and male-identified participants have significantly different ratings on sloppy *himself* trials. We argue that this motivates more careful reporting of participant demographic information when binary gender is treated as an independent variable in the design of trial items.

1. Introduction

In this paper, we examine the behaviour of English object reflexives under VP ellipsis, as exemplified in (1):

(1) Josh$_i$ saw himself$_i$ and Tom did too

At issue is the interpretation of the elided VP in the second conjunct: is the reading sloppy, where Tom reflexively sees himself, or is the reading strict, where Tom also sees Josh?

This is by no means a new issue. Hestvik (1995) claims that it is "uncontroversial that reflexives have sloppy readings" (p. 212), while noting that there is considerable disagreement in the literature on the status of strict readings. He argues that the availability of strict readings depends on the structure of the VP ellipsis in question. Hestvik's key claim is that strict readings under coordination, as in (1), are difficult for speakers to obtain, while subordination facilitates strict readings:

(2) Hannah voted for herself because Rose did

In (2), Hestvik reports that it is much easier for speakers to get a strict reading of the reflexive *herself*, understanding Rose to have voted for Hannah.

In his overview of the issues, Hestvik also cites gender mismatch effects noted in Kitagawa (1991), which uses (3) as a point of departure:

*We would like to thank the reviewers and audience at the CLA for their comments on this work. This research was funded by URGC SSHRC Explore Faculty Seed Grant 1050868 to Storoshenko.

Actes du congrès annuel de l'Association canadienne de linguistique 2022.
Proceedings of the 2022 annual conference of the Canadian Linguistic Association.
© 2022 Dennis Ryan Storoshenko and Jesse Weir

(3) John considers himself to be intelligent, and Mary does too
(Kitagawa 1991 p. 519)

According to Kitagawa, some consultants allow both sloppy and strict readings for (3), while others report only the strict; the gender mismatch apparently makes the sloppy reading unavailable for some consultants. In contrast, for a parallel example where *Bill* appears in place of *Mary*, obviating the gender mismatch, the sloppy reading is universally accepted, with only some consultants accepting the strict. The facts change for (4):

(4) Mary considers herself to be intelligent, and John does too
(Kitagawa 1991 p. 520)

In this case, Kitagawa reports universal rejection of the sloppy reading, regardless of how individual consultants had judged (3). This leads to the conclusion that while some English ideolects can tolerate a mismatch for *himself* under ellipsis, none can interpret *herself* with a non-feminine binder. This claim is the one we test in our study.

Using a contextual felicity task, we find that while Kitagawa is correct in noting an asymmetry between *herself* and *himself*, the effect is more subtle and complex than he describes. We find no evidence for gender mismatching facilitating strict readings in coordinated VP ellipsis, but we do find a difference in judgements of *herself* and *himself* in sloppy conditions, with participants' gender identity being more predictive of results than the (mis)matching ϕ-features in the items themselves.

This work is preliminary. In addition to not yet considering possible interactions when considering coordination versus subordination, our discussion is limited to the binary-gendered forms *herself* and *himself*, mirrored in our participant pool which is limited to participants identifying as either female or male. The reasons for this are twofold. First, acknowledging that there is a gender-neutral form of the third person reflexive available in English, evidence suggests that inter-speaker variation in acceptance of *themselves* versus *themself* exists (Ackerman, 2018; Conrod et al., 2021), and including either might add an extraneous variable to our study. Secondly, with trial items limited to binary gender, a parallel limitation in the gender identity of the participant pool allows for the most rigorously factorialized design, and is most likely to expose potential interactions. Furthermore, this is a detail not explored in Kitagawa's description of the facts, which we believe to be an obvious factor to check. Without a doubt, the results of this study suggest that more diverse trial items and participant pools will be an illuminating path for future work.

The rest of the paper is structured as follows. In Section 2, we update the theoretical discussion of deriving reflexives, with emphasis on the proposals in Kratzer (2009), and the predictions they make with regards to Kitagawa's examples. Section 3 takes a brief look at psycholinguistic literature on the interaction between gendered antecedents and English reflexives, noting some gaps in existing work. Our study design and results are presented in Section 4. A discussion of the results follows in Section 5, putting the results of the present study into context via comparison with prior results reported in 2017, and with items testing reflexives in manner *by*-phrases. Section 6 outlines potential future work.

2. Sloppy Ellipsis and Binding

The works cited above lay out some expected loci of variation in terms of the availability of sloppy readings for reflexives, however they use quite different theoretical frameworks. Kitagawa's ideolects are essentially parametric variations in Principle A of the Chomsky (1981) Binding Theory, while Hestvik's discussion is based in Discourse Representation Theory. The basis for our study's predictions will be a more updated approach to reflexives, that of Kratzer (2009), following the implementation described in Johnson (2014). To illustrate, we present Johnson's analysis of a simple reflexive sentence as in (5):

(5) Tom sees himself

According to Kratzer, the initial step of the derivation for this sentence would be the merge of the verb *sees* (abstracting away from the details of tense and subject verb agreement) with a minimal pronoun *n*:

(6) $[_{VP}$ sees n]

At this stage, *n* is a featureless object, and following the treatment of *v* developed in (Kratzer, 1996), the VP is a saturated one-place predicate. The next merge is with a *v* head acting as a reflexivizer. In addition to introducing a merge position for an external argument, this *v* head λ-binds the minimal pronoun, yielding the desired reflexive semantics. As shown in (7), this *v* head also carries a bundle of features, in this case [3.SG.MASC,REFL]:

(7) $[_{v'}\ v_{[3.\mathrm{SG.MASC,REFL}]}\ [_{VP}$ sees n]]

In binding the *n*, the features of *v* unify with the features of *n*, yielding the final spellout form *himself*. Finishing the derivation of *v*P, *Tom* merges at the specifier position:

(8) $[_{vP}$ Tom $[_{v'}\ v_{[3.\mathrm{SG.MASC,REFL}]}\ [_{VP}$ sees n]]]

At this point, there is again a unification between features of the specifier *Tom* and the *v* head, in this case ensuring that the ϕ-features match. With this in place, the morphosyntax and semantics of the reflexive are set, and the derivation proceeds as normal. A desirable consequence of this analysis is that with a lexicon containing an appropriate collection of reflexivizing *v* heads, it is impossible to derive a sentence such as (9):

(9) * Tom saw myself

The derivation for this sentence would crash at the point of feature unification between the *v*P specifier and the *v* head. Also assuming that these *v* heads are the only route to reflexive semantics, Principle A of the Binding Theory no longer needs to be independently stipulated; Principle A effects are rendered epiphenomenal in this analysis.

The derivation of sloppy readings under VP ellipsis is straightforward under this account; as the VP's theme underlyingly has no ϕ-features, a similarly featureless VP in the elided conjunct would predict that feature mismatches such as those in both (3) and

(4) should be equally well-tolerated. Under this analysis, the type of gender contrast described by Kitagawa would only emerge if the minimal pronoun *n* inherently contained a [FEM] feature in (4), but no equivalent [MASC] feature existed in (3). Kratzer does allow for the possibility that English minimal pronouns can be base-generated with a [PL] to account for some number mismatches, but there is no equivalent discussion of gender.[1] If the effect Kitagawa describes is found, then the inventory of minimal pronouns would need to be expanded. An asymmetry between gender features is however not predicted by contemporary analyses of English ϕ-features, as both Bjorkman (2017) and Konnelly and Cowper (2020) propose models of English features in which there is no markedness contrast between [FEM] and [MASC]. Rather, each gender feature stands in equal contrast to an unmarked third person value which spells out the *they/them* paradigm.

3. Gender Effects

The question of gender mismatches under discussion here intersects with the recent discussion of gender biases in linguistic publishing. Kotek et al. (2021) draws attention to a trend of favouring masculine subjects in linguistic example sentences. If this trend held through literature on VP ellipsis, it could conceivably be the case that the effect Kitagawa observes would go unnoticed, as the type of example sentence necessary to uncover the contrast would simply never appear. Even in the original Kitagawa paper, the gender contrast is not part of the primary discussion, and all examples of sloppy identity are constructed with masculine names as apparent default, while strict identity examples tend to be more diverse.

While Kotek et al. (2021) covers theoretical literature, making recommendations for the use of introspective and corpus data, it is relatively silent on the issue of experimental and psycholinguistic work. Here, not only should researchers be conscious of the examples used in the discussion of their research, there is also a largely open question of whether bias in the construction of trial items should be controlled for. In the case of sloppy ellipsis, the Hestvik characterization of the issue as “uncontroversial” suggests that any interference of gender effects would be unexpected, though Kitagawa notes a sharp contrast.

Another reason to expect that reflexive trial item content might influence study results is that reflexive binding is often used in studies on the nature of gender stereotypes in English. One example is in Kreiner et al. (2008), where the research question differentiates between profession nouns that have stereotypical gender associations (e.g. *general*, *secretary*) versus what they call definitional gender associations in nouns such as *king* or *princess*. Using a (mis)matching reflexive paradigm as in (10), they test whether the two different types of gender information are used differently in reference resolution.

(10) Adapted Kreiner et al. (2008) paradigm, male associations

a. Yesterday the minister left London after reminding himself/herself about the letter

[1] This uncertainty regarding a plural feature is another reason to step back from the discussion of *themself/selves*, as it is unclear which feature bundle is responsible for each form of the reflexive.

b. Yesterday the king left London after reminding himself/herself about the letter

Also testing positional variables in anaphoric versus cataphoric contexts, they find that the two types of gender information are processed differently. This builds on earlier work such as Garnham et al. (2002) which tests whether subverting binary gender stereotypes using descriptors of clothing or other accessories (e.g. ties versus skirts) yields different results than subversions arising from so-called "biological" factors such as giving birth or having a mustache, or Carreiras et al. (1996) which tests simple gendered pronoun binding.

One weakness of many of these studies, which are framed on an assumed gender binary reflected as an independent variable in the trial items, is that the gender identity of the participants is not only not treated as a variable to be considered in analysis, it is rarely even reported.[2] One might expect that perception and intensity of gender stereotype are closely connected with gender identity, and so more detailed reporting is called for. Indeed, one study of profession nouns in English, Kennison and Trofe (2003), finds that the gender identity of participants interacts with the intensity of stereotypes. Addressing this methodological question — whether participant gender identity interacts with an assumed binary in trial items — is part of the motivation for the controlled participant pool in our study. While the binding effect that Kitagawa describes is not as grounded in stereotypes as the work on profession nouns, it could be the case that the asymmetry in sloppy ellipsis also has roots in a bias regarding the relative flexibility of gender roles. If the facts are as categorical as described, then the distinction might be best described in terms of different morphosyntactic features; if effects are more subtle, then the explanation may lie outside the grammar. With these questions in mind, we turn in the next section to our study.

4. Contextual Felicity Study

To investigate the availability of strict and sloppy readings under coordinated VP ellipsis, we conducted a web-based contextual felicity study, recruiting a balanced sample of binary gender-identified participants. This strategy overcomes what we believe to be a weakness of work described in the prior section, failing to consider that binarized gender in trial items might interact with binarized gender among participants. We first discuss the design and trial items of our study, followed by a brief discussion of participant recruitment. The section concludes with a discussion of the results and statistical analysis.

4.1 Design

Our study uses a text-based contextual felicity task to collect judgements on the availability of different readings of sentences containing coordinated VP ellipsis. In each trial, participants read a text describing a scenario compatible either with a sloppy reading (two

[2] None of the three studies listed gives gender identity information for participants.

individuals in parallel reflexive events) or a strict reading (the first conjunct understood reflexively, and the subject of the second conjunct also acting upon the subject of the first). We illustrate this with one of our trial items, as in (11):

(11) Kara forced herself into the elevator, and Marissa did too

Two different frame contexts are constructed for this elevator scenario example:

(12) a. **Strict** Kara and Marissa were trying to get on a very crowded elevator. Kara pushed her way into the elevator, and left room for Marissa, who was getting off first. When Marissa got off, she pushed back on Kara's shoulder to make sure the door would still close.

b. **Sloppy** Kara and Marissa were trying to get on a very crowded elevator. Kara pushed her way into the elevator. Still seeing room, Marissa also squeezed into the elevator. It was an uncomfortable elevator ride, but everyone got to their floors.

With these frame contexts, the first independent variable, **Reading**, is set with two levels, Strict or Sloppy. For each scenario, four different trial sentences are constructed. Again, using the elevator scenario:

(13) a. Joel forced himself into the elevator, and Martin did too

b. Joel forced himself into the elevator, and Marissa did too

c. Kara forced herself into the elevator, and Marissa did too

d. Kara forced herself into the elevator, and Martin did too

With this paradigm, we see our other two independent variables: **Reflexive** with two levels *himself* and *herself*, and **Match** with two levels Match, where the gender of the name in the second sentential conjunct matches that of the first, and Mismatch, where the second conjunct subject has the opposing binary gender. The selection of proper names for this study is based on an examination of birth records from the Province of Alberta in the year 2000. For this and other projects on English, we have identified the top 100 most popular entries appearing only on the list of recorded girl names, and the top 100 appearing only on the list of recorded boy names. The choice of this method of name identification is based on the fact that most studies in our lab are conducted using local undergraduate populations; choosing frequent birth names from the local community at roughly the same birth year of our mean participant age should accurately reflect the biases of our typical participant pool without a need for norming.

To ensure that there are no extraneous gender biases in our examples arising from the predicates, we first conducted a separate study norming study where participants were presented with two sentences as in (14) as opposing poles of a 7-point Likert scale:

(14) a. The man forced himself into the elevator

b. The woman forced himself into the elevator

Participants would be presented with both sentences in (14), placed, for example, such that (14a) is a 1 on the scale, and (14b) is a 7. Participants were instructed to rate the relative acceptability of each sentence, such that a selection of 4 indicates equivalent acceptability, while moving toward one or the other end of the scale indicates that the sentence at that end of the scale is more acceptable. Using some control pairs with one clearly grammatical and one clearly ungrammatical sentence, we verified that participants understood the task. With 20 participants recruited from the University of Calgary community (10 female-identifying and 10 male-identifying), we tested 100 different predicates in pairings similar to (14), using contrasts such as *the man/the woman* or *the boy/the girl*. To be included in this study, no predicate's mean rating could be significantly different from 4, the neutral midpoint, and among those meeting this criterion, we selected those with means closest to 4. With predicates and names chosen this way, we constructed 48 different scenarios, each with two context frames with proper names adjusted to fit the four different trial item variations. This yields a 2x2x2 within-participants design for our study, with eight trial item lists presented as a Latin-square. Each participant saw six items in each condition, with each scenario presented only once.

On each trial, participants were first presented with the context sentence, needing to press the space bar to bring up the sentence to be rated, and a visual reminder of the rating scale. Felicity ratings were on a 7-point Likert scale, with 1 indicating that the sentence is not compatible with the context, and 7 indicating that the sentence is compatible with the context. Before the presentation of test trials, participants were presented with three practice trials followed by on-screen feedback, making it clear that ratings are not expected just to be for acceptability, but for logical coherence with the context provided. In addition to the 48 trial items, participants saw 52 filler trials, all coordinated VP ellipsis sentences, some including reflexives in manner *by*-phrases, *one*-anaphora, and some testing unrelated issues of scope interactions with ellipsis. Embedded within the 52 filler trials were eight items where the target sentence is unambiguously false relative to the given context. All participants saw all eight of these in exactly the same contexts; if any participant rated four or more of these items at 4 or higher on the contextual felicity scale, their data was removed from the study, and they were replaced in recruitment.

4.2 Participants and Data Removal

Our study was conducted online, implemented using PsychoPy 3 (Pierce et al., 2019) on the Pavlovia platform. Participants were recruited using Prolific Academic, paid at a rate of £7.50/hour. Using the demographic filters within the system, we selected only native English speakers, initially recruiting 24 male-identified participants, and 24 female-identified participants. We did not independently ask for participant gender identity, relying solely on the information provided to Prolific. As participants were removed from the study based on their performance on the unambiguous filler items, new ones were recruited to maintain the balanced sample. In doing so, our experimental design is now raised to 2x2x2x2, with a between-participants variable of gender identity.

72 participants were recruited, with data from 24 participants removed. Removed participants had more than four errors on the unambiguous trials, though this was correlated with participants completing all 100 trials (plus practice and debriefing) in six minutes or less. As data collection proceeded, participants whose overall study duration was less than six minutes were automatically removed, as we find it implausible that participants could get through 20 trials per minute given the amount of text they needed to read. We attribute the high level of data removal to the long, reading-intensive nature of the study, as many participants showed signs of fatigue, reaching a point where they were just button-pressing to reach the end. The excess recruitment was required in order to obtain 24 usable data files for female-identified participants, and 24 for male-identified. Among the 48 submissions analyzed, ages range from 18 to 56 years old, with a mean of 28.65 for female-identified and 27.29 for male-identified participants. Country of origin was not controlled.

After removing uncooperative participants, we used a second metric to remove outlier responses from the remaining 48 participants' data. Taking the sum of the time from the first display of a context to the reaction time to the trial sentence, we calculated a mean trial time of 22.74s, with a standard deviation of 21.17s. Any trials taking over 90s were removed as being likely points where participants' attention shifted, and any taking less than 4s were removed, as indications that participants had likely not read the context completely. This removed 2.9% of the total responses from the remaining 48 participants. With this, the mean trial time drops to 21.78s, but the standard deviation becomes 13.15s, suggesting the data are much more representative of cooperative participants overall.

4.3 Results

Mean ratings by condition are presented in Figure 1. Data are analyzed in R (R Core Team, 2017), using the lme4 and lmerTest packages (Bates et al., 2015; Kuznetsova et al., 2017). For the analysis of rating data, we used a subtractive method of linear mixed effect model comparison, beginning with the most complex interaction model, as in (15). After inspection of the resulting model, the most insignificant interaction is removed to create a new model, with the two models compared using the `anova` function from lmerTest. If the models are not significantly different, the model with the lowest AIC value (typically the simpler model) is taken to be superior. Iteratively removing interactions and fixed factors, we arrive at an interim fixed effect model, to which we then begin adding random slope factors, to arrive at the best-fitting characterization of by-participant and by-item variation. The final best-fitting model for these data is shown in (16). The summary of fixed effects for this model is shown in Table 1.

(15) Starting Model
Rating˜Reflexive*Reading*Match*Gender + (1|Participant) + (1|Item)

(16) Best-Fitting Model
Rating˜Reflexive*Reading*Gender + (1+Reading|Participant) +
(1+Gender|Item)

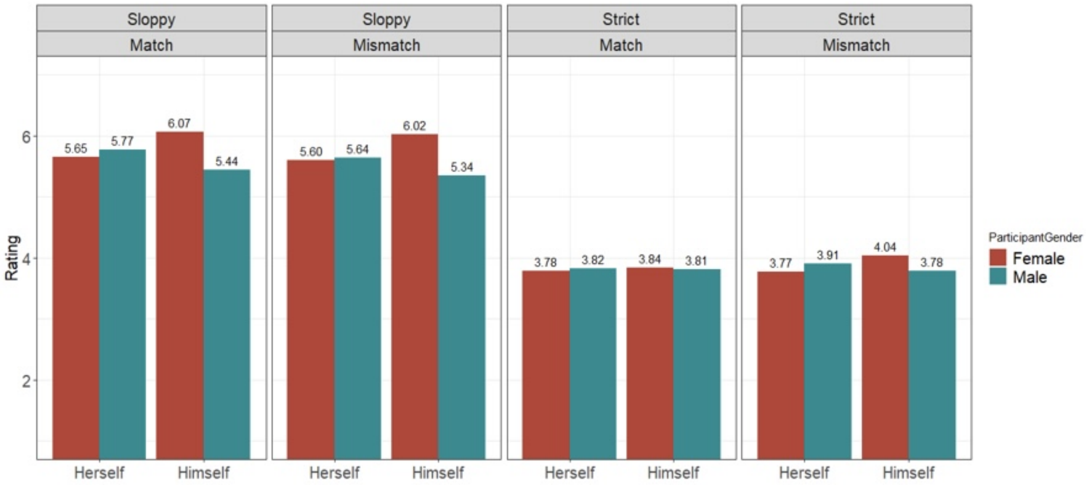

Figure 1. Mean Ratings for all Conditions in Present Study

	Estimate	Std Error	df	t-value	Pr(> |t|)	
Intercept	5.6127	0.1681	101.982	33.386	<2e-16	***
Reflexive *himself*	0.4114	0.1685	363.621	2.411	0.0151	*
Gender Male	0.0839	0.2080	63.128	.404	0.6879	
Reading Strict	-1.8281	0.2652	85.123	-6.891	9.02e-10	***
Reflexive:Gender	-0.7157	0.1735	371.259	-4.124	4.61e-05	***
Reflexive:Reading	-0.2412	0.2388	364.084	-1.010	0.3132	
Gender:Reading	-0.0021	0.3372	56.964	-0.006	0.9950	
Reflexive:Gender:Reading	0.4704	0.2456	368.024	1.916	0.0562	.

Table 1. Fixed Effect Summary for Model in (16)

The most striking difference between the starting model and the best fitting model is that the Match factor has been simplified out. This is not surprising in light of the picture painted in Figure 1: within both the Sloppy (left) and Strict (right) pairings, there is little difference between the Match and Mismatch conditions. There is virtually no difference at all on the Sloppy side, suggesting that mismatched ϕ-features are easily tolerated, and while there is a slight improvement for the Mismatch-*himself* condition, it is negligible and thus it is not surprising that this factor falls away from the model. The fixed effect of Reading is quite clear, as the Sloppy conditions are overall more accepted than Strict, though the Strict conditions are perhaps not as degraded as might be expected, with all means hovering near 4 on the 7-point scale.

The other two factors remaining in the model, Reflexive and Gender, are most clearly reflected in the Sloppy conditions. Again referring to Figure 1, we see that in the *herself* trials, there is virtually no difference in the mean ratings of the female- and male-identified participants. However, for the Sloppy-*himself* conditions, there is a clear effect where the female-identified participants rate these higher than the *herself* items, while the male-identified participants rate these as lower than the *herself* items. This interaction appears in both the Sloppy-Match and Sloppy-Mismatch conditions.

The random slopes included in the best-fitting model indicate that across participants, some are more sensitive to the Strict/Sloppy contrast than others, and some items get more divergent readings between female- and male-identified participants than others. To further isolate the participant gender effect, we conducted a series of planned pairwise comparisons, running t-tests between the ratings from female- versus male-identified participants within each of the eight conditions defined by the between-participant variables. As might be predicted upon inspection of Figure 1, only two of these emerge as significant to the $p < 0.01$ level: the Sloppy-Match-*himself* condition, and the Sloppy-Mismatch-*himself* condition. While there is a slight mirroring of this preference for *himself* among female-identified participants in the Strict-Mismatch condition, it does not approach significance. As the effects of participant gender identity seem limited to the Sloppy conditions, we did one last comparison pooling all of the Sloppy condition data together (ignoring the apparently irrelevant factor of Match), checking whether there was a difference between the female- and male-identified participants in their ratings of *herself* versus *himself* trials. For the male-identified participants, there was no significant effect. For the female-identified participants, there is some evidence of a preference for *himself* trials over *herself* trials, but this is only with a p-value of 0.03, which is not significant once appropriate Bonferroni corrections have been applied.

5. General Discussion

Based on these results, Hestvik's claim that that sloppy readings for reflexives under coordinate VP ellipsis are uncontroversial is clearly supported. Across the board, sloppy readings are well-accepted, with condition means all at approximately 5.75 on the 7-point scale. These are somewhat lower than expected, as there are examples of filler items being regularly rated over 6. To put these ratings in context, it is worth comparing these results with unpublished results presented at the CLA in 2017 (Storoshenko, 2017). In that presentation, results on a contextual felicity task with items similar to the ones in the present study were shown, as in Figure 2, though it should be noted that the studies differed in a number of crucial ways.

First of all, the 2017 reflexive items were more limited in number than in the 2022 study, as anaphoric items were used as distractors in the 2017 study whose main focus was on scope and ellipsis. The reflexive section of the study was based on a simpler 2x2 design contrasting Strict versus Sloppy readings again, and testing for (Mis)matching ϕ-features. Participants only saw three items in each of the four conditions. Because of this small number of items, the reflexive form *herself/himself* could not be reliably treated as a within-participants variable. Furthermore, due to an error in stimulus design for that study, more *himself* trials were presented than *herself* trials, offsetting attempts to counterbalance this variable across the participant pool. Furthermore, gender-identity was not controlled within that participant pool (n=40), with the vast majority self-identifying as female. The

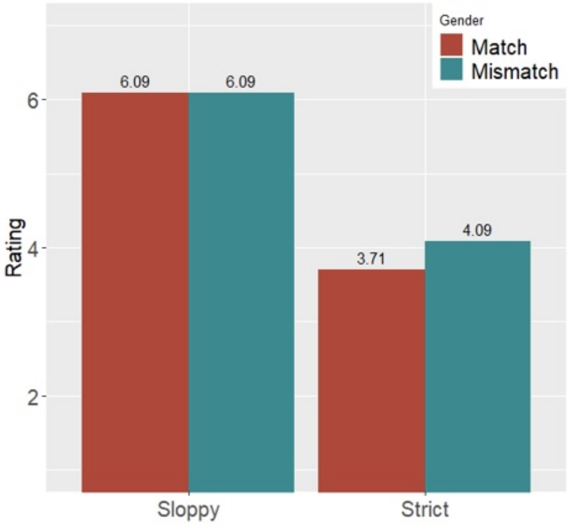

Figure 2. Mean Ratings for all Reflexive Conditions in the 2017 Study

present study was designed to improve upon the 2017 one by quadrupling the number of reflexive items, allowing for control over the form of the reflexive, while also strategically recruiting to explore the untested question of the effect of participant gender identity. Data for the 2017 study were collected in-person at the University of Calgary.

The relatively high acceptance of the Strict conditions in the present study is not an isolated result, as similar results were obtained in the 2017 lab study. In light of the proposed binding analysis, strict readings of these trial items should be impossible, yet both studies found them to be rated at the middle of the 7 point scale, in contrast to other items, to be discussed below, which received much lower ratings. We believe it to be unlikely that these unexpectedly high ratings are the result of the reflexive pronouns being interpreted using a different mechanism, such as co-reference, due to the lack of a (mis)match effect. If participants were able to treat *self*-pronouns as parallel to ordinary referential pronouns, we expect they would be merged whole-cloth, inherently containing [FEM] or [MASC] feautres, and therefore more susceptible to the manipulation of gender features in the trial items. Instead, it seems that participants parse the trial items as uniformly containing an underlyingly featureless minimal pronoun in the object position, and are accommodating a form they recognize to be marked in the given context. This accommodation is reflected in a separate analysis of trial response time, where Strict conditions are slower.

Another effect from 2017 replicated in the present study is the insignificance of the Match factor. This supports the Kratzer-inspired analysis of English reflexives, refuting the contrast proposed by Kitagawa. Not only do we find no evidence for a special treatment of *herself*, if anything, the results of our study point to the uniqueness of *himself*. The locus of this distinction is in our between-participants variable though, and therefore not likely to be captured by proposing a difference in ϕ-features for the underlying minimal pronoun.

Overall, the results from 2017 effectively foreshadow those of the present study.

There is no significant effect of the Match factor, and Sloppy conditions are rated significantly higher than Strict ones. One difference is that the means are slightly higher in the 2017 study, with the Sloppy conditions both averaging over 6 on the scale, while Strict conditions are rated roughly the same as in the present study. There are two possible explanations for this distinction. One is that it is a task effect, reflecting the greater distraction inherent in a study conducted online versus in the lab, with an overall larger number of items. A second possibility is that this is a manifestation of the difference between female- and male-identified participants in our present study. Extrapolating from the results in Figure 1, we might expect that the mean rating for Sloppy items in a study with a participant pool that is mostly female-identifying, and with a set of items containing more *himself* than *herself* items, would be closer to 6 on the 7-point scale. We cannot concretely decide between these two possibilities, as both the relatively lower overall condition means and the preference for *himself* among female-identified participants are limited to the Sloppy conditions of the present study.

To attempt to tease apart the issue if whether or not the relatively lower ratings for the Sloppy conditions are a result of running the present study online, we will briefly examine another set of similar items tested in both the present study and the more limited 2017 version. These involve sentences with a reflexive pronoun in a manner adjunct *by*-phrase, rather than in the object position of a sentence with coordinated VP ellipsis, as in (17):

(17) Andy wrote a proposal by himself, and Colin did too

For this type of sentence, it is quite easy to construct a Sloppy context: we simply tell a story in which Andy writes a proposal with no assistance, and then Colin also writes a proposal with no assistance. However, there is no real “Strict” interpretation, akin to Colin writing a proposal with help from Andy. For this reason, instead of the Reading variable here being Sloppy versus Strict, it is Sloppy versus what we describe as Non-Parallel, a context in which Andy writes a proposal without assistance, but Colin has help. In this way, these trials act as another set of control items. According to our intuitions, the Sloppy reading is easily accessible, while the Non-Parallel readings are false, expected to get low ratings. In the 2017 study, these were implemented again with the limited 2x2 design, again with three of each condition. For the present study, this was implemented with the full 2x2x2x2 design, though each participant saw only two trials per condition. In Figures 3 and 4, we present the results from the 2017 study and the present study, respectively, simplified down to the same 2x2 variables of Reading and Match.

In both the 2017 data and the present study, we find a significant effect of Reading, and no effect or interaction with Match. However, we again observe a drop in acceptability across the board when comparing the 2017 results to the present ones: in Figure 3 both Sloppy conditions are well above 6, while the present study in Figure 4 has both below 6, though somewhat higher than some of the co-argument reflexive Sloppy conditions. This suggests that the difference between the older lab-based task versus the online study has had some impact on the ratings. As for the other independent variables introduced in the present study, there is no fixed effect of participant gender-identity in the manner adjuncts, but there

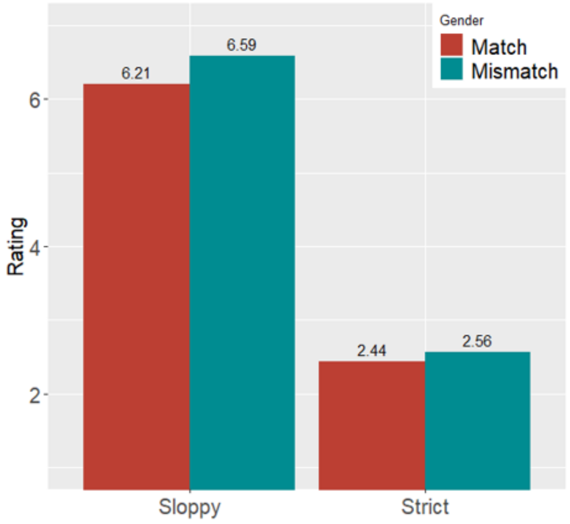

Figure 3. Mean Ratings for Manner Adjunct Conditions in the 2017 Study

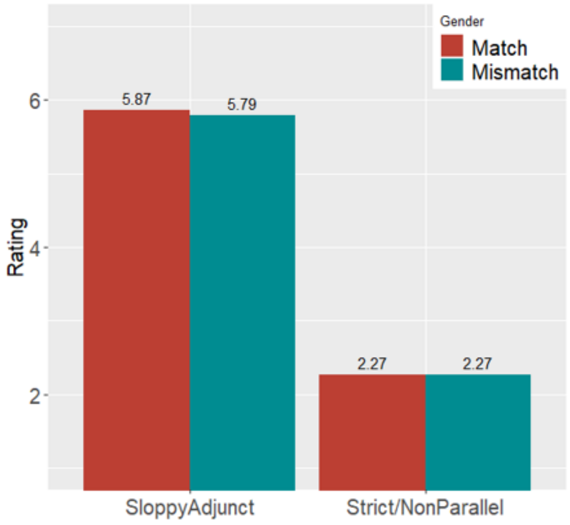

Figure 4. Mean Ratings for Manner Adjunct Conditions in the Present Study

is again a marginally significant (p=0.056) interaction whereby the Sloppy condition items are rated higher with *himself* than *herself*. While we have no explanation for this effect, we note that the same thing is observed for the co-argument reflexives, and suggests a similar mechanism is at play in considering these items.

We must finally highlight the finding that has emerged from treating participant gender identity as a controlled variable. Limited only to co-argument reflexives under coordinated VP ellipsis, we find a clear pattern where female-identified participants rate Sloppy-*himself* trials higher than Sloppy-*herself* trials, while male-identified participants rate Sloppy-*himself* trials lower than Sloppy-*herself* trials. While this difference of about

0.7 on our 7-point scale is not enough to change the overall interpretation of results comparing the Sloppy versus Strict conditions, there are instances in the literature where such minute distinctions are taken to be meaningful when statistical testing shows significance. While it is clear that participant gender identity does not interact with the Match factor — the initial concern that led us to control this variable — a subtler interaction with the Reflexive form variable has emerged.

6. Conclusion and Future Work

In this study, we have demonstrated that under coordinated VP ellipsis, English reflexives are indeed interpreted sloppily. Further, we have shown that mismatched gender features in the third person do not degrade the acceptability of sloppy readings, nor do mismatches make strict readings more available. Instead, we see that strict interpretations of the reflexive are significantly less acceptable than sloppy ones, but not as categorically unacceptable as plainly false sentences, as in the manner adjunct fillers. This supports Johnson's move in applying Kratzer's derivation of bound variables to these cases.

The results of this study also give a hint as to the derivation of reflexive manner *by*-phrases in English. With the clear tolerance of mismatched ϕ-features, and an easily accessible sloppy interpretation now verified through experimental testing, an argument that the reflexive inside the adjunct is derived in a parallel manner to the co-argument reflexive can be made. This leaves the door open to a full formal analysis, considering whether the relevant binding is also from a *v* head checking against the merged external argument, or from a different position higher up the clausal structure accessing the subject's position after EPP movement. Answering these questions would illuminate the position of the manner adjunct within *v*P, and give some suggestions as to its formal semantic characterization.

We have also observed an unexpected interaction between participant gender identity and the form of the reflexive, but only in the co-argument position. The origin of this interaction remains unknown, though we believe this points to a need for, at the very least, clear reporting of participant demographics in studies where gender ϕ-features are part of the independent variables in study design. Done consistently across different study paradigms, a pattern may yet emerge leading to an explanation. While we have no reason at this point to believe that this interaction is a result of anything in the grammar, its limitation only to co-argument positions is puzzling. This apparent limitation could simply be a result of the fact that there are far fewer data points in our examination of the manner adjuncts, and the effect is too subtle to emerge from the available data. One possible next step would be to replicate this part of the study with more items, and to this we would consider also adding subordinate ellipsis. As Hestvik (1995) notes, the facts around the availability of sloppy readings are different in cases such as (2), and it might be interesting to see whether the same interaction with participant gender identity emerges in a similarly-controlled study. As stated at the outset, expansion beyond the gender binary, both in terms of trial items and especially in the participant pool, is clearly called for, now that there is evidence that the between-participants variable makes a difference.

References

Ackerman, Lauren. 2018. Being themself: Processing and resolution of singular (im)personal *they*. Presented at the 31st CUNY Conference on Human Sentence Processing.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software* 67(1): 1–48.

Bjorkman, Bronwyn M. 2017. Singular *they* and the syntactic representation of gender in English. *Glossa* 2(1): 80.

Carreiras, Manuel, Alan Garnham, Jane Oakhill, and Kate Cain. 1996. The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology* 49A(3): 639–663.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Conrod, Kirby, Byron Ahn, and Ruth Schultz. 2021. How many *selves* for *them*? Poster presented at the 2021 Northeast Linguistic Society (NELS) Conference.

Garnham, Alan, Jane Oakhill, and David Reynolds. 2002. Are inferences from stereotyped role names to characters' gender made elaboratively? *Memory and Cognition* 30(3): 439–446.

Hestvik, Arlid. 1995. Reflexives and ellipsis. *Natural Language Semantics* 3: 211–237.

Johnson, Kyle. 2014. Commentary on 'Gender mismatches under nominal ellipsis' by Jason Merchant. *Lingua* 151: 33–42.

Kennison, Sheila M., and Jessie L. Trofe. 2003. Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycolinguistic Research* 32(3): 355–378.

Kitagawa, Yoshihisa. 1991. Copying identity. *Natural Language and Linguistic Theory* 9(3): 497–536.

Konnelly, Lex, and Elizabeth Cowper. 2020. Gender diversity and morphosyntax: An account of singular *they*. *Glossa* 5(1): 1–19.

Kotek, Hadas, Rikker Dockum, Sarah Babinski, and Christopher Geissler. 2021. Gender bias and stereotype in linguistic example sentences. *Language* 97: 653–677.

Kratzer, Angelika. 1996. Severing the external argument from its verb. In *Phrase structure and the lexicon*, ed. Johan Rooryck and Laurie Zaring, 109–138. Dordrecht: Kluwer Academic Publishers.

Kratzer, Angelika. 2009. Making a pronoun: Fake indexicals as windows into the properties of pronouns. *Linguistic Inquiry* 40(2): 187–237.

Kreiner, Hamutal, Patrick Sturt, and Simon Garrod. 2008. Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language* 58: 239–261.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H.B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13): 1–28.

Pierce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindelov. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51(1): 195–203.

R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Storoshenko, Dennis Ryan. 2017. Mismatched gender in elided *self*-pronouns and *one*-anaphora. Poster presented at the 2017 Annual Conference of the Canadian Linguistic Association (CLA), Toronto, ON.