More than One Agreement Planning Mechanism: Reflexive Number Source Sensitive to Syntactic Structures

Zhenghao Herbert Zhou (herbert.zhou@yale.edu)
Department of Linguistics, Yale University

Yuyang Liu (yuyang.liu.yl2472@yale.edu)
Department of Linguistics, Yale University

Abstract

Prior studies on agreement attraction in sentence production have investigated the differences in verb and reflexive agreement planning, as evidenced by a distinction in their error rates. This study investigated two possible accounts: Planning Order, where verbs and reflexives share the same underlying mechanism that is applied at different stages, and Number Source, where verbs and reflexives retrieve features from different sources. To further adjudicate between the two, we extended the empirical ground to reflexives in complement and adjunct positions. Results replicated previous findings on the distinction between verbs and reflexives, while no significant difference was found in attraction effects among reflexive conditions, thereby supporting Number Source. Predicted by neither account, we novelly observed a significant increase in baseline error rate for adjunct but an insignificant increase for complement, suggesting a finer-grained Number Source model sensitive to syntactic structural differences.

Keywords: sentence production; agreement attraction; reflexive agreement; scene-description paradigm

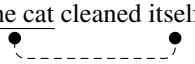
Introduction

Fluent speech requires speakers to plan upcoming materials while simultaneously articulating prepared materials, and how this planning unfolds over time remains a central question in sentence production. One way to investigate the underlying mechanisms of sentence production is by examining error rates in producing syntactic or semantic dependencies, among which *agreement* (Chomsky, 2000, 2001 et seq.) is one of the most prominent cases.

Existing psycholinguistic studies (e.g.: Dillon et al., 2013; Kandel et al., 2022; Wagers et al., 2009) have mainly focused on two types of agreement patterns: (i) verb agreement as in (1a), where the inflection of the main verb matches the features of its subject, such as person and number in English; and (ii) reflexive agreement as in (1b), where a reflexive pronoun (characterized by *-self* or *-selves* in English) matches the person, gender, and number of the antecedent to which it refers. Asterisks represent ungrammaticality.

(1) a. *Verb agreement*
The cat is/*are sleeping.

b. *Reflexive agreement*
The cat cleaned itself/*themselves.



This study concerns number agreement, whose errors in sentence production are likely to occur at the presence of an intervening constituent bearing a different number feature (Dillon et al., 2013; Wagers et al., 2009). Such errors are demonstrated in (2), where the verb/reflexive mistakenly agrees with the intervening constituent and thus fails to agree with its subject/antecedent, a phenomenon known as *attraction*. In both examples below, we bracket the attractor and underline the actual, intended antecedent or goal of agreement.

(2) a. *Verb agreement attraction*
*The cat next to [the dog] are sleeping.

b. *Reflexive agreement attraction*
*The cat next to [the dogs] cleaned themselves.

If the two types of agreement follow exactly the same planning procedure, their error rates will share a similar pattern, but this is not the case. A series of studies in sentence production (Kandel & Phillips, 2022) shows that verb agreement has a significantly higher error rate than reflexive agreement, as in Table 1 (where S = singular, P = plural, and the first/second letter marks the number of the subject head/attractor), and the same asymmetry has been reported in sentence comprehension (Dillon et al., 2013). This distinction motivates further study of how this distinction arises, as well as the extent to which the two types of agreement still share a common part of the underlying mechanism and the extent to which they differ.

Table 1: Percentage of agreement errors in Kandel and Phillips's (2022) Experiments 1–2.

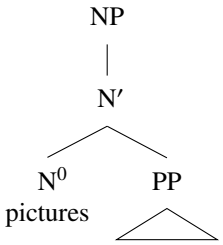
Condition	Verb (%)	Reflexive (%)
SS	0.4	0.2
SP	24.8	0.4
PP	3.0	1.1
PS	21.2	3.3

Against this background, we advance existing approaches by addressing a key gap in previous studies, which have restricted their attention to reflexives in coargument positions (here, the object position). We extend the experimental setup to reflexives contained in prepositional phrases (PPs) and further distinguish between PPs functioning as complements (headed by *of*, etc.) and those functioning as adjuncts (headed by *next to*, etc.). Within an X-bar theoretic framework (Chomsky,

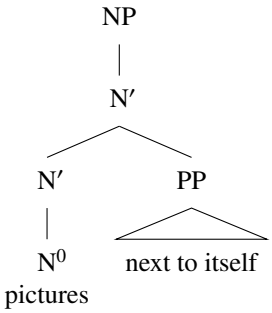
In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

1970 et seq.) for instance, the complement-adjunct distinction is reflected in their merging sites: complement PPs merge with the head, whereas adjunct PPs merge with an intermediate bar projection. As such, adjuncts are merged (at least) one node higher in the syntactic structure than complements. We illustrate this structural contrast in (3–4), using *pictures of itself* and *pictures next to itself* as examples. NP stands for *noun phrase*.

(3) *Complement*



(4) *Adjunct*



By extending the investigation to reflexives in complement and adjunct PPs, we are able to further evaluate two competing hypotheses for the source of the distinction in the error rate between verb and reflexive agreements, which previous work was unable to differentiate clearly. The following section outlines the two hypotheses considered in this study.

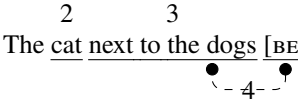
Hypotheses

Planning Order

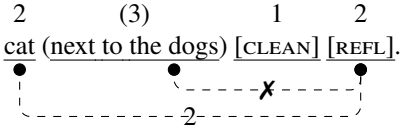
The Planning Order Hypothesis posits a single retrieval mechanism that accesses number features from planned materials, shared by both verb and reflexive agreements but applied at different planning stages (Kandel & Phillips, 2022). (See also Momma and Ferreira, 2021.) Verb agreement is planned after the encoding of the full subject NP, including the attractor, whereas reflexive agreement is planned no earlier than the subject head but before the subject modifier (i.e., the attractor). In this framework, the chance of retrieving the wrong number feature from the attractor is significantly lowered in reflexive agreement, since at the time of its planning, the attractor has not yet been planned, making its features unavailable to access. By contrast, retrieval misfire remains possible for verb agreement because at the time of planning the inflection of the verb, the attractor has already been encoded and obtains features accessible for agreement, thus explaining its much higher error rate.

To illustrate, take (2) as examples. Under the Planning Order Hypothesis, materials already planned at the time of planning the number feature of the elements undergoing agreement are presented in (5), where a parenthesized position represents that the corresponding material has not yet been planned, and REFL stands for *reflexive*.

(5) a. *What has been planned in (2a)*



b. *What has been planned in (2b)*



In (5b), the subject (head) and the reflexive refer to the same entity. Their tight semantic coupling (e.g.: Ahn, 2014; van Hoek, 1997) motivates the planning of them with close temporal proximity, crucially before the attractor, as signaled by the numbers above the example, which stand for the order of planning. As a result, when the reflexive decides its number feature, it can only retrieve the singular feature from the planned subject, avoiding a retrieval misfire from the attractor, which has not yet been planned. On the contrary, in (5a), although the verb lemma planning may precede the subject, the inflection of the verb, lacking the strong semantic relation with the subject as in reflexive agreement, is proposed to be planned late, after both the subject head and the subject modifier. Hence, when the verb computes its number, it can incorrectly retrieve the number feature from the attractor.

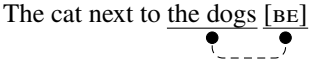
In sum, the Planning Order Hypothesis assumes one single underlying mechanism for both agreement types. However, this mechanism is applied at different planning stages for different types of linguistic materials, causing the distinction in error rates between verb and reflexive agreements.

Number Source

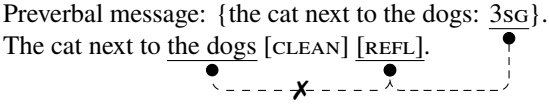
The Number Source Hypothesis leverages on a distinction in the source of the number feature and posits two routes of retrieval: a "linguistic" one and a "conceptual" one (Eberhard et al., 2005). (See also Kandel et al., 2024, 2025 among others.) While verb agreement still retrieves the number feature from previously planned linguistic materials as shared by the Planning Order Hypothesis (through the "linguistic route"), reflexive agreement retrieves the number feature directly from the conceptual representation in the preverbal message (through the "conceptual" route). The conceptual route makes reflexives intrinsically robust to agreement errors, since the the conceptual source directly interacts with semantic interpretation and is free from syntactic attraction.

Consider (2) again. We illustrate the number sources of the two agreement types in (6).

(6) a. *Number source in (2a)*



b. *Number source in (2b)*



In (6a), verb agreement retrieves the number feature from previously planned linguistic materials, where the presence of *the dogs* can induce attraction, just as under the Planning Order Hypothesis. As for reflexive agreement in (6b), instead of retrieving the number feature from linguistic materials like *cat* or *dogs*, it retrieves features directly from the conceptual representation of the subject, given that the conceptual representation of both the subject and the object are linked to the same entity.

To summarize, the Number Source Hypothesis claims that there are two distinct underlying mechanisms for verb and reflexive agreements with respect to the source of feature retrieval. Verb agreement retrieves the number feature from the morphosyntactic representations of the linguistic materials and is susceptible to attraction, whereas reflexives retrieve the number feature directly from the conceptual representation and is robust to attraction.

Predictions in the Current Study

Although the two hypotheses make the same predictions on verb and coargument reflexive agreements, they would diverge on reflexives in noncoargument conditions. When the reflexive is not a coargument with its antecedent, the PP in which the reflexive resides, be it noun complement or adjunct, may be planned after the heads of the arguments. This timing opens up the possibility of the reflexive to be planned no earlier than the attractor, providing possibilities of attraction errors.

Therefore, the Planning Order Hypothesis predicts a higher error rate of reflexive agreement in noncoargument conditions than the coargument one. On the other hand, the Number Source Hypothesis contends that a noncoargument reflexive will retrieve the number feature directly from the preverbal message without interference, just as in the coargument condition. Therefore, the reflexive agreement error rates across all conditions should not be significantly different from each other. Results in the current study favor the Number Source Hypothesis, as discussed in the upcoming sections.

Methods

Experimental Paradigm

To elicit the relatively natural sentence production processes, we adapted the *scene-description paradigm* from Kandel and Phillips 2022, in which participants viewed pictures without written prompts and produced sentences describing the depicted event. This choice was motivated by the goal of eliciting the full planning processes — from formulating a preverbal message to incrementally mapping that message onto linguistic structures — rather than having participants memorize and complete a provided preamble (Bock & Miller, 1991), which would involve different mental processes.

To support naturalistic production while tightly controlling the intended structures, the study embedded events in an alien world. Participants learned three alien types (*blueys*, *greenies*, and *pinkies*) and an action called *mimming*. In a mimming event, the actor showed "pulses," and the undergoer had its antenna light up. The alien world also included pictures that could be described from the subject alien's perspective as either *a picture of itself/themselves* (when the picture depicts aliens) or *a picture next to itself/themselves* (when the picture depicts a scenic spot). To reliably elicit an intervening NP between the subject (agreement *controller*) and the element undergoing agreement (agreement *target*), each trial presented two panels with different layouts, where only one panel depicted the target mimming event. This disambiguation pressure encouraged productions like "The blueys above the greeny . . . " so that the description uniquely identified the intended subject group, thereby placing a local NP intervener between the subject and the agreement target. See Figure 1 as an example.

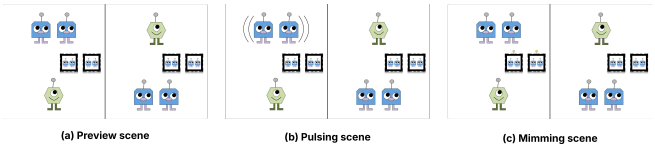


Figure 1: Example stimulus for "The blueys above the greeny mimmed the pictures of themselves."

Materials

Conditions The critical sentence conditions consist of VERB (verb agreement) and three reflexive agreement conditions: (i) COARGUMENT, where the reflexive is a core argument of the predicate; plus two noncoargument reflexive constructions where the reflexive appears inside the object NP, (ii) COMPLEMENT (reflexive inside a complement PP of the picture NP) and (iii) ADJUNCT (reflexive inside an adjunct PP modifier of the picture NP). See Table 2 for example sentences. Across all conditions, subject and intervener numbers were fully crossed to create four number combinations. For COMPLEMENT and ADJUNCT, we set the object (*picture(s)*) number to always agree with the subject number.

To reduce strategic expectations that every transitive trial requires a reflexive (which could distort message planning), the experiment also included two filler types: (i) the object is the intervener NP, thereby eliciting *it/them* (e.g., "The blueys above the greeny mimmed it.") and (ii) the object is a third alien type distinct from both subject and intervener, thereby eliciting a full NP rather than any pronoun (e.g., "The blueys above the greeny mimmed the pinkies.")

Stimuli Construction Stimuli were cartoon scenes constructed from a small, fully crossed inventory. Each NP position had 6 realizations (3 alien types × singular/plural), and the intervener relation used one of two prepositions (*above* or *below*). The intervener alien type always differed from the subject. All target sentences used the same nonce verb (*mim*). In total, the experiment comprised 336 trials. To keep participation time manageable (under 45 minutes per participant), items were distributed across six lists. Each list contained 8

Table 2: Sample sentences of the critical conditions in plural-singular.

Condition	Sample Sentence
COARGUMENT	The blueys above the greeny mimmed {themselves, *itself}.
COMPLEMENT	The blueys above the greeny mimmed the pictures of {themselves, *itself}.
ADJUNCT	The blueys above the greeny mimmed the pictures next to {themselves, *itself}.
VERB	The blueys above the greeny {were, *was} mimming.

verb trials and 48 reflexive trials (8 per reflexive condition and 24 filler trials), with the order pseudorandomized.

Participants

The study was conducted with the approval of the authors' affiliation's Institutional Review Board. Participants (N = 112) were recruited on Prolific.[1] Eligibility required native English, holding a Bachelor's or higher degree, age between 18 and 60, normal vision/hearing/speaking ability, and prior Prolific experience (at least three completed studies). Informed consent was obtained from all participants. Responses from 88 participants were retained for analysis after exclusion (criteria described below).

Procedure

Stimuli Presentation The experiment was implemented in PennController for IBEX (PCIbex; Zehr and Schwarz, 2018) and completed on the participants' laptops via Prolific. Participants were assigned to one of six stimulus lists. Each stimulus was presented in size 1920 × 1080 pixels, centered on a white background. Speech was recorded using built-in microphones (44.1 kHz by default).

One Single Trial A single trial presented a short scene sequence followed by production and two post-production tasks. Verb trials used two scenes, whereas reflexive trials used three. The preview and an additional pulsing scene (for reflexive trials to indicate the actor) were shown for 1500 ms each, followed by the target mimming scene for 3000 ms, with automatic transitions. Participants then pressed a button to initiate speaking from an instruction screen with the scenes no longer visible, heard a beep marking recording onset, and produced the description with no time limit but instructions to speak naturally. Recording ended when participants clicked a button labeled *Done*. After production, participants first went through an attention check — a binary choice task checking the location of a cross briefly flashed for 500 ms — which was intended to impose mild cognitive load. Then, they completed a binary true-or-false choice comprehension question about the described scene, judging a statement about the actor or undergoer of the mimming event they just described. False statements were constructed by replacing either the type or number of the questioned actor/undergoer.

[1]Prolific. (2024). *Prolific* [Online participant recruitment platform]. https://www.prolific.com

Structure of Trials The study had two main blocks. In the beginning of Block 1, participants were introduced to the alien world and the intransitive use of *mimming*. They were explicitly instructed to disambiguate descriptions with {*above*, *below*} NPs. They then went through two familiarization trials, with intended sentences displayed on the screen as feedback. After familiarization, 8 VERB trials were presented without filler. Block 2 introduced the picture setting in the alien world and subsequently the transitive use of *mim* for eliciting the reflexive and filler structures. Participants practiced across five reflexive conditions (each of the three experimental conditions and the two filler conditions) with the intended target sentence as feedback on the screen, followed by 48 experimental trials. They were encouraged to take a short break after finishing the first half of the trials. Recordings were uploaded to the Bluehost server after completion.[2]

Data Processing Audio and postproduction task data were obtained from 112 participants, and all recordings were transcribed using OpenAI's Whisper `turbo` model (Radford et al., 2023).[3] Each participant conducted 8 trials for every condition (3584 total data points). Each transcription was then labeled via a two-step pipeline. First, a conservative automated Python heuristic examined the produced sentence structure, extracted the relevant number features, and assigned a preliminary label (agreement error vs. no error vs. `CheckAgain` for undecidable cases). Second, experimenters manually checked the audio, verified the transcription/target, and assigned a final outcome label. A transcription would be labeled *invalid* if it failed to follow the intended sentence structure (e.g., by not including the intervening NP) or failed to produce the relevant number (e.g., using the past tense in VERB or not producing reflexive pronouns in the reflexive conditions).

We then applied data exclusion at the participant-by-condition level following these criteria: (i) if comprehension question accuracy was lower than 75%, all data were excluded; (ii) if more than 8 trials across the 32 experimental trials were labeled invalid, all data were excluded; and (iii) if a participant produced at least 4 invalid trials (out of 8) in a given condition, all 8 trials from that participant in that condition were excluded. After these exclusions, 2672 data points (74.5%) from 88 participants were retained for downstream analyses.

[2]https://www.bluehost.com

[3]Whisper is a transformer-based speech recognition model trained on large scales of data. It is able to recognize and transcribe nonce words like *bluey* and *mimming*.

Results

Model Fitting

To investigate the error distribution patterns, we fit generalized linear mixed-effects models (GLMM) with a binomial distribution and logit link using the package `lme4` v1.1.37 (Bates et al., 2015) in R v4.3.1.[4] We first fit one model for each of the four conditions in order to assess attraction patterns, where the model contained fixed effects of **match** (with SS/PP coded as `match` and SP/PS as `mismatch`, dummy-coded) and **subject number** (`singular` and `plural`, dummy-coded), as well as their interaction. The model also contained a random intercept for participants and a random slope of match by participant to account for variability across individuals. Since the stimuli were fully crossed and balanced, we did not expect an effect on specific words or sentences. We thus did not include a random effect of item. The formula for the model was:

(7) error ~ match * subject_num + (1 + match | participant)

We then fit a model that compared all four conditions. It contained an additional mixed effect of **sentence type** (the four conditions, dummy-coded) and its interactions with match and subject number. This specification directly tested whether construction types modulate overall error rates and whether attraction-like mismatch effects differ across constructions and antecedent number.

To further interpret pairwise contrasts across sentence conditions, we computed estimated marginal means (EMMs) in the GLMM models with the `emmeans` v1.10.1 package (Lenth & Piaskowski, 2026). We used `tukey` adjustment to control the family-wise error rate when conducting pairwise contrasts between factor levels, and we converted logit predictions back to probabilities for better interpretability.

Findings

Overall Error Rates across Conditions Figure 2 shows the overall error rates of the four conditions, aggregating over the four number combinations. Consistent with previous literature, Coargument yielded the lowest error rate (20/689 = 2.90%), significantly lower than Verb (53/614 = 8.63%). In between, Complement had a higher error rate (29/699 = 4.15%) than Coargument, and Adjunct had an even higher error rate (48/670 = 7.16%). The error rate of Complement was not significantly higher than that of Coargument (β = 0.299, SE = 0.377, p = 0.427), while Adjunct (β = 1.229, SE = 0.325, $p < 0.001$) and Verb (β = 1.053, SE = 0.402, $p < 0.01$) had significantly higher error rates than Coargument.

Significant Attraction Effect for Verb but not for Reflexives In order to investigate the agreement attraction patterns, we broke down each condition into the four number combinations, as shown in Figure 3. Showing agreement attraction

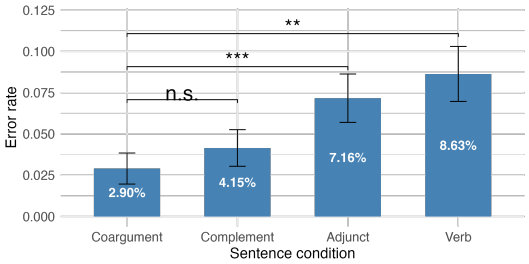


Figure 2: Error rates by conditions.

means having a significant match effect, where `mismatch` is more likely to lead to errors than `match`, represented by a positive β value. GLMM models on individual sentence conditions showed that Verb showed a significant match effect (β = 5.747, SE = 2.304, p = 0.013), where the estimated error probabilities from EMM were 0.75% (SE = 0.48%) in match conditions and 8.71% (SE = 2.06%) in mismatch conditions. On the other hand, for all three reflexive conditions, although there was a tendency of attraction (match $\beta > 0$), none of them was significant. For Coargument (β = 1.003, SE = 2.913, p = 0.731), the estimated error probabilities were 0.65% (SE = 0.37%) in match conditions and 1.37% (SE = 0.57%) in mismatch conditions. For Complement (β = 1.628, SE = 2.754, p = 0.554), the estimated error probabilities were 0.69% (SE = 0.39%) in match conditions and 2.32% (SE = 0.83%) in mismatch conditions. As for Adjunct (β = 0.252, SE = 0.814, p = 0.757), the estimated error probabilities were 2.22% (SE = 0.94%) in match conditions and 4.35% (SE = 1.24%) in mismatch conditions. Although the estimated error probabilities for mismatch conditions were consistently higher than for match conditions for all reflexive conditions, thereby showing an attraction-prone pattern, none of them was significant.

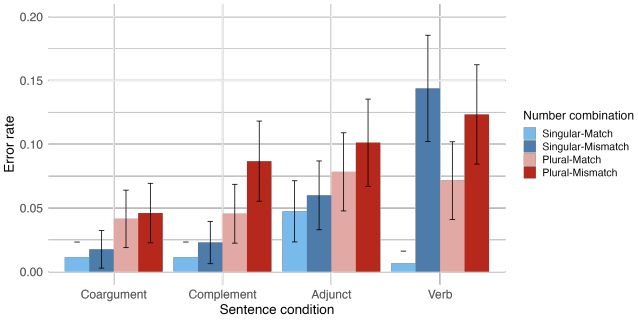


Figure 3: Error rates by conditions and number combinations.

Markedness Asymmetry in Reflexives We also investigated which subject number is more likely to elicit attraction errors. Known as the markesness asymmetry, previous studies have found that singular subjects are more likely to elicit attraction errors than plural subjects in verb agreement (e.g., Wagers et al., 2009), and plural subjects are more likely

[4] R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org

than singular subjects in reflexive agreement (e.g., Kandel & Phillips, 2022). In our models, a positive β for subject number means plural subjects have higher error rates than singular ones. For reflexives in COARGUMENT and COMPLEMENT, a significant subject number effect was found (subject number $\beta > 0$). For COARGUMENT ($\beta = 2.378$, SE = 1.145, $p = 0.038$), the estimated error probabilities were 0.50% (SE = 0.27%) in singular subjects and 1.77% (SE = 0.65%) in plural subjects. For COMPLEMENT ($\beta = 2.355$, SE = 1.052, $p = 0.025$), the estimated error probabilities were 0.58% (SE = 0.30%) in singular subjects and 2.74% (SE = 0.91%) in plural subjects. As for ADJUNCT ($\beta = 0.631$, SE = 0.493, $p = 0.201$), the estimated error probabilities were 2.23% (SE = 0.78%) in singular subjects and 4.34% (SE = 1.28%) in plural subjects. This indicates that for reflexive production, plural subjects are more likely to elicit errors than singular subjects, consistent with previous findings.

Discussion

To summarize: (i) we replicated the distinction on agreement attraction profiles between VERB and COARGUMENT using the same scene-description paradigm as in Kandel and Phillips 2022; (ii) we found attraction-prone though insignificant error distribution patterns for all three reflexive conditions, such that mismatch conditions are (insignificantly) more likely to elicit agreement errors than match conditions; (iii) our post-hoc analyses suggested that the difference in attraction effects among the three reflexive conditions was insignificant. While it is possible that we were underpowered to detect potentially significant differences in the attraction effects across the three reflexive conditions (as predicted by Planning Order), since no reliable difference in attraction effects is observed across the three reflexive conditions, our current results favor Number Source over Planning Order.

As side findings: (iv) we observed significant differences in the absolute attraction rates across four conditions, such that COARGUMENT $\approx$ (n.s.) COMPLEMENT $<$ ADJUNCT $\leq$ VERB; (v) we observed robust markedness asymmetry for all three reflexive conditions, such that in contrast to verb agreement, plural subjects are significantly more likely to elicit agreement errors than singular subjects in reflexive agreement. We discuss potential reasons and implications of our findings below.

Comparing Attraction Rates with Kandel and Phillips 2022 The error rates we elicited in the current experiment diverge from those reported by Kandel and Phillips (2022): we have lower error rates for VERB and higher error rates for COARGUMENT. (See Table 1 for comparison.) We suspect the following reasons: in the current experiment, each participant only did 8 VERB trials, compared to 96 trials in Kandel and Phillips 2022. The additional cognitive load imposed by the large number of trials might lead to more attraction errors. On the other hand, for reflexive conditions, Kandel and Phillips (2022) had 192 trials per participant that involved producing a COARGUMENT or simple pronoun sentence. Although participants only did 48 trials in the current study, they varied across five sentence constructions, with the two non-coargument cases more complex than the others. Therefore, additional reflexive errors in the current study could be explained by the extra cognitive load needed to distinguish the varied target constructions.

Explaining Differences in Absolute Error Rates across Reflexives Although no significant attraction effect was found in the three reflexive conditions, we observed that the absolute error rate of ADJUNCT was reliably higher than COARGUMENT, as well as an insignificant difference between COMPLEMENT and COARGUMENT. Differing the significant distinction between ADJUNCT and COMPLEMENT to the next point, we suspect the following reasons for the differences: COMPLEMENT and ADJUNCT have more complex sentence structures and lexical items than COARGUMENT, which complicated the planning process. Moreover, they also have one additional linear intervener (the object *picture(s)*) than in COARGUMENT, which, at least according to Planning Order, imposes additional cognitive load in the reflexive planning process. Even though we designed the stimuli to always let the object share the same number as the subject, thereby decreasing the probability of making attraction errors, this might still lead to an increase in baseline error rates. Future studies could consider varying the object number to increase the chance of attraction to see whether an even higher baseline error rate would be elicited.

Further Specification of Number Source As discussed in findings (ii–iii), we found no significant differences of the attraction effects across the reflexive conditions, supporting Number Source. Nonetheless, as in finding (iv), we unexpectedly observed that the reflexive conditions differ in absolute error rates, and the difference is especially significant between ADJUNCT and the others. To the best of our knowledge, this finding is a novel and has not been the focus of the literature, nor is it explained by either of the two accounts for attraction mechanisms: Number Source predicts no difference across the reflexive conditions, and Planning Order predicts differences in attraction effect but *not* baseline error rates.

The differences in baseline error rates between the two non-coargument conditions suggest further specification of the Number Source Hypothesis such that the planning mechanism of reflexive agreement must be sensitive to syntactic structures. Given that accessing conceptual representations could be susceptible to errors (as demonstrated by the nonzero error rates), the specification might lie in the timing or manner in which adjuncts access conceptual representations, one that is different from heads or complements. This distinction would stem from the difference in the height of merging sites in syntax (e.g., (3–4)) between complements and adjuncts and makes adjunct reflexive agreement more error-prone than the other conditions. Our conclusion is consistent with recent proposals by Kandel et al. (2024) that both the conceptual route and the linguistic route are involved in pronoun production. Taking together, our results characterized further details of Number Source for future studies to investigate.

Acknowledgments

We express our gratitude to members of Yale's Language and Brain Lab, the anonymous HSP and CogSci reviewers. We also thank Suhas Arehalli, Brian Dillon, Robert Frank, Margaret Kandel, Shota Momma, and Michael Wilson for helpful discussion and feedback. All errors are our own.

In this study, we used GPT-4o, ChatGPT 5.2, and Microsoft Copilot for: (i) help with generating the starter Python scripts that generate the full picture stimuli from primitive components, generating the full set of experimental and filler sentences, and pseudorandomizing item orders in every list; (ii) help with generating code snippets for PCIbex experimental interface design; and (iii) help with generating the R scripts for data analysis, model fitting, and plotting. All scripts were further checked and modified by the experimenters, and all outputs have been manually checked. No AI tools were used for writing the manuscript.

References

Ahn, B. (2014). *Giving reflexivity a voice: Twin reflexives in English* [Doctoral dissertation, University of California].

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*. https://doi.org/10.18637/jss.v067.i01

Bock, K., & Miller, C. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93. https://doi.org/10.1016/0010-0285(91)90003-7

Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in English transformational grammar* (pp. 184–221). Ginn.

Chomsky, N. (2000). Minimalist inquiries: The framework. In R. Martin, D. Michaels, J. Uriagereka, & H. Lasnik (Eds.), *Step by step: Essays on minimalist syntax in honor of Howard Lasnik* (pp. 89–156). MIT Press.

Chomsky, N. (2001). Derivation by phase. In M. Kenstowicz (Ed.), *Ken Hale: A life in language* (pp. 1–52). MIT Press. https://doi.org/10.7551/mitpress/4056.003.0004

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. https://doi.org/10.1016/j.jml.2013.04.003

Eberhard, K., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531–559. https://doi.org/10.1037/0033-295X.112.3.531

Kandel, M., Pañeda, C., Bahmanian, N., Martinez Bruera, M., Phillips, C., & Lago, S. (2025). Number and grammatical gender attraction in Spanish pronouns: Evidence for a syntactic route to their features. *Journal of Cognition*, *8*, Article 10. https://doi.org/10.5334/joc.416

Kandel, M., & Phillips, C. (2022). Number attraction in verb and anaphor production. *Journal of Memory and Language*, *127*, Article 104370. https://doi.org/10.1016/j.jml.2022.104370

Kandel, M., Wyatt, C., & Phillips, C. (2022). Agreement attraction error and timing profiles in continuous speech. *Glossa Psycholinguistics*, *1*, Article 4. https://doi.org/10.5070/G601157

Kandel, M., Wyatt, C., & Phillips, C. (2024). Number attraction in pronoun production. *Open Mind*, *8*, 1247–1290. https://doi.org/10.1162/opmi_a_00167

Lenth, R., & Piaskowski, J. (2026). *emmeans: Estimated marginal means, aka Least-squares means* [R package version 2.0.3]. https://rvlenth.github.io/emmeans

Momma, S., & Ferreira, V. (2021). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, *128*, Article 101397. https://doi.org/10.1016/j.cogpsych.2021.101397

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518.

van Hoek, K. (1997). *Anaphora and conceptual structure*. University of Chicago Press.

Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. https://doi.org/10.1016/j.jml.2009.04.002

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832