GLOSSA psycholinguistics

Sturt, P. 2022. Syntactic and semantic mismatches in English number agreement. *Glossa Psycholinguistics* 1(1): 19, pp. 1–33. DOI: https://doi.org/10.5070/G6011159

eScholarship
University of California

Syntactic and semantic mismatches in English number agreement

Patrick Sturt, University of Edinburgh, UK, patrick.sturt@ed.ac.uk

In English, it is possible for a morphologically singular collective noun like *government* to control both singular (syntactic) agreement and plural (semantic) agreement in the same sentence (e.g. *The government has praised themselves*). It has been claimed that sentences with the opposite pattern of agreeing elements are ungrammatical (e.g. **The government have praised itself*), and there is a corresponding asymmetry in corpus frequencies of these two configurations. Across two acceptability judgement experiments, we show that the acceptability contrast is affected by the relative order of the two agreeing elements, with degraded acceptability in the case where the first agreeing element shows plural agreement and the second shows singular agreement, relative to the opposite configuration. This pattern is found both when the agreeing verb precedes the reflexive, and when the reflexive precedes the verb. Overall, the results suggest that the initial formation of a semantic agreement dependency between an agreement target and a collective controller makes subsequent morpho-syntactic agreement with the same controller less accessible. We argue that any theoretical account of these results would require an important role for incremental processing.

Glossa Psycholinguistics is a peer-reviewed open access journal published by the eScholarship Publishing. © 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

OPEN ACCESS

1. Introduction

Agreement has been a major focus of research in psycholinguistics, because it can illumiate how linguistic structure interacts with other aspects of cognition, such as memory access in comprehension (e.g. Dillon et al., 2013; Pearlmutter et al., 1999; Wagers et al., 2009) or planning processes in production (e.g. Bock et al., 1999; Bock et al., 2004; Eberhard et al., 2005; Franck et al., 2006; Haskell & MacDonald, 2003).

Agreement has been characterised as "the systematic covariance between a semantic or formal property of one element and a formal property of another" (Steele; 1978, p. 610), highlighting the fact that it can also be seen as part of the interface between syntax and interpretation. In this paper, we report two studies on English number agreement that investigate this interface. In the sentence *The boys laugh*, the (plural) number feature of the subject (*the boys*) matches with that of the verb (*laugh*). Plurality is a formal property both of the verb *laugh* and of the subject *the boys*, because these two elements are both morphosyntactically plural. It is also a semantic property of *the boys*, because this phrase refers to more than one individual. For nouns like *boy/boys*, the formal and semantic number properties align: if the noun is morphosyntactically singular, its referent is also semantically singular, and if it is morphosyntactically plural, its referent is also semantically plural. However, for collective nouns like *committee*, there can be a lack of alignment between morphosyntactic and semantic number: the noun *committee* is formally singular, but it may denote an assemblage of multiple individuals, making the referent semantically plural. If such a noun participates in subject-verb agreement, or anaphor-antecedent agreement, then the agreeing element may match either the semantic or the syntactic feature, as shown in (1):[1]

(1) a. The committee was featured in a TV documentary.
b. The committee were featured in a TV documentary.
c. The committee filmed itself for a TV documentary.
d. The committee filmed themselves for a TV documentary.

In (1a,c), the agreeing element (*was*, *itself*) takes the singular form, agreeing with the morphosyntactic features of *committee*, while in (1b,d), plural agreement is used. We will refer to the former type of agreement as *syntactic* agreement and the latter as *semantic* agreement. We use these terms following convention,[2] although in the case of collective nouns, the singular agreement option could be also be argued to reflect the semantic feature, namely that *committee* denotes a single group. Note, however, that the value of this semantic feature correlates with morpho-syntax: multiple groups require plural marking (*committees*). In contrast, the fact that a

[1] However, see Bock et al. (2006) and Levin (2001) for relevant differences between American and British English.

[2] It should be pointed out that some theoretical accounts make a slightly different distinction (e.g. Ackema & Neeleman, 2018, make a distinction between *restrictor agreeement* vs. *syntactic agreement*, which does not coincide exactly with the conventional distinction between semantic and syntactic agreement).

committee is composed of multiple individuals is not reflected in the morphology of the noun, justifying the use of the term "semantic agreement" for cases where plural elements agree with *committee* in terms of this feature. In general, semantic agreement is widely attested cross-linguistically, and it can affect a variety of different linguistic features, including but not limited to number and gender (Corbett, 1979, 1983, 2000).

According to the Agreement Hierarchy (Corbett, 1979), the distribution of semantic agreement varies by dependency type cross-linguistically, with elements that are structurally closer to the agreement controller being more biased towards participating in syntactic agreement (e.g. determiner-noun agreement), and elements more structurally distant to the controller having a relatively greater possibility for semantic agreement (e.g. anaphor-antecedent or pronoun-antecedent agreement). Moreover, the hierarchy is intended to be implicational, such that if a given dependency type allows semantic agreement for a particular language, then all dependency types that are further towards the semantic side of the hierarchy will also allow semantic agreement in that language. The agreement hierarchy is illustrated in **Table 1**, along with examples involving English number agreement:

Table 1: The agreement hierarchy as it applies to English number agreement. The row named "Dependency type label" shows the category names used in the Agreement Hierarchy literature. Cross-linguistically, semantic agreement is more likely as one moves towards the right of the hierarchy. The "English example" row shows examples of number agreement dependencies in English that exemplify the relevant categories. Note that there is a "Relative Pronoun" category on the hierarchy, but this is left empty in the Table, because English relative pronouns are not marked for number.

	←SYNTACTIC			**SEMANTIC→**
Dependency type label	**Attributive**	**Predicate**	**Rel. Pronoun**	**Personal Pronoun**
English example	Determiner-noun	Verb-subject	—	Anaphor-antecedent
Syntactic agreement:	This group	The group agrees	—	The group filmed itself.
Semantic agreement:	*These group	The group agree	—	The group filmed themselves.

As shown in **Table 1**, determiner-noun agreement is an instance of an Attributive dependency type, and therefore occupies the syntactic end of the hierarchy. This dependency type does not allow semantic agreement in English (i.e. *these committee* is ungrammatical). In contrast, verb-subject and anaphor-antecedent dependencies, being instances of Predicate and Personal Pronoun dependencies respectively, do allow semantic agreement, although psycholinguistic and corpus evidence suggests that there are processing and distributional differences between these two types of dependencies, aligning with their positions on the hierarchy. One source of evidence for this comes from elicited speech production, where studies have shown that collective noun controllers lead to higher rates of plural agreement for anaphor-antecedent dependencies than they do for verb-subject dependencies (Bock et al., 2004; Bock et al., 1999). A similar asymmetry has been found in naturally occurring corpus data (Levin, 2001), and overall, the direction of this effect is what would be expected on the basis of the agreement hierarchy, with the higher rates of semantic (i.e. plural) agreement for the anaphor-antecedent dependencies aligning with these dependencies appearing further towards the semantic end of the hierarchy, relative to verb-subject dependencies. A phenomenon that may be related to this is the fact that quantificational phrases like *everyone* can antecede plural anaphors (e.g. *Everyone praised themselves*), while plural agreement on the verb is not possible, even where plural anaphor agreement is present (i.e. *Everyone were praising themselves* is ungrammatical).

As well as the production evidence discussed by Bock et al. (1999) and Bock et al. (2004), there is also comprehension-based evidence suggestive of the idea that the processing of verb-subject and anaphor-antecedent dependencies aligns with the agreement hierarchy. In a series of eye-tracking experiments reported by Kreiner et al. (2013), there was evidence for temporary processing difficulty for a plural marked verb in the context of a subject headed by a collective noun (e.g. *the family… want*), relative to appropriate control conditions (Kreiner et al., 2013, Experiments 2 and 3). However, no such evidence was evident for plural anaphors referring to a singular collective noun phrase (e.g. *the family … themselves*), even though reference to a non-collective singular subject did cause statistically detectable difficulty (e.g. *The schoolgirl … themselves*). Although this evidence is somewhat circumstantial, the findings suggest consistency with the agreement hierarchy, given that there was evidence for verb-subject dependencies to be sensitive to morpho-syntactic agreement, but the evidence was not found for anaphor-antecedent dependencies.

Bock et al. (1999) and Bock et al. (2004) suggest that the effects that they found in their production studies can be explained in terms of a sequential model of language production (see also Levelt, 1989), whereby the process of selecting a singular or plural prounoun occurs at an earlier stage of production where conceptual information (e.g. about the semantic plurality of collectives) is available, while the process of appending an agreement suffix on the verb occurs at a later stage, where grammatical information is more relevant. However, the results of Kreiner et al. (2013) suggest that the biases of the agreement hierarchy may instead be a more general feature of both production and comprehension (see Kreiner et al., 2013, for a further discussion of this point).

In this paper, we consider cases where the same agreement controller participates in two number agreement dependencies within the same sentence, as in (2) below, where *the government* participates in one dependency with *avoid/avoids*, and another one with *themselves/itself*:

(2) a. The government constantly avoid criticizing themselves about the unbalanced budget.
b. The government constantly avoids criticizing itself about the unbalanced budget.
c. *The government constantly avoid criticizing itself about the unbalanced budget.
d. The government constantly avoids criticizing themselves about the unbalanced budget.

In (2a,b), both the anaphor and the verb match in number—in (2a) both show plural (i.e. semantic) agreement, and in (2b), both show singular (i.e. syntactic) agreement. However, in (2c,d) these two elements mismatch in number. It has been claimed that there is an acceptability contrast between examples (2c) and (2d). Examples like (2c), where the verb shows plural agreement and the anaphor shows singular agreement are claimed to be less acceptable than (2d), which has the reverse configuration (Huddleston & Pullum, 2002; P. Smith, 2015, 2017), although some authors claim that both (2c) and (2d) are ungrammatical, as we will discuss in the General Discussion. Examples analogous to (2c) are also less frequently attested in corpora than those analogous to (2d) (Levin, 2001). In this paper, we will refer to the pattern of acceptability illustrated in (2a–d) as the “double mismatch effect”.

To our knowledge, the only study that has investigated double mismatches with collective nouns using experimental methods is Kreiner et al. (2013). Experiment 3 of Kreiner et al. (2013) is an eye-tracking experiment that used sentences like those in (2) among other control conditions. However, analysis of eye-movement measures in the reflexive and following region did not show significant effects indicating a mismatch cost between verb and reflexive number, whether the verb was singular and the reflexive was plural (2d), or the reverse configuration (2c). On the other hand, an ERP study reported by Molinaro et al. (2008) showed evidence of processing difficulty for a number agreement mismatch between verb and reflexive, although their study did not use collective nouns. They used sentences like (3a–d):

(3) a. The famous dancer was nervously preparing herself to face the crowd.
b. The famous dancer were nervously preparing themselves to face the crowd.
c. The famous dancer were nervously preparing herself to face the crowd.
d. The famous dancer was nervously preparing themselves to face the crowd.

In ERPs measured at the reflexive (*herself/themselves*), Molinaro et al. (2008) reported a P600 effect for conditions (3c) and (3d), where the reflexive and verb mismatched in number, relative to the fully grammatical and matching control condition (3a). This suggests, contrary to the null result of Kreiner et al. (2013), that mismatches between reflexive and verb number can indeed elict measurable processing difficulty, at least given certain experimental methods. However, given that Molinaro et al.’s study did not use collective nouns, it is not fully relevant to the present paper. Moreover, although Molinaro et al. (2008) counterbalanced their experimental stimuli among those with a singular and a plural subject, they did not report separate analyses for the two resulting mismatch patterns (i.e. those where the verb is singular and the reflexive is plural, vs. those in the opposite configuration).

To summarize the discussion above, an intuitive acceptability contrast has been proposed between two different patterns of mismatch between verb and anaphor number in relation to a collective noun phrase (see (2c,d) above). Experimental evidence for this asymmetry has not been provided to date, although a related study has shown that mismatches between anaphor and verb number in general can lead to detectable processing difficulty (Molinaro et al., 2008) with non-collective nouns. In the following paragraphs, we consider two different ways of stating the descriptive generalization to capture the double mismatch effect. One way of stating this generalization is in terms of the positions of the two agreement targets on the agreement hierarchy:

> In addition to the sentences where the agreements match, we also expect that a mismatch between the two targets can arise if it is the element to the right on the hierarchy that shows semantic agreement, and the element to the left that shows morphological agreement. (P. Smith, 2015, p. 203)

According to this form of generalization, the relative acceptability of (2d) is due to the fact that the verb *avoids* shows morpho-syntactic agreement, and is further towards the left (i.e. syntactic agreement) side of the agreement hierarchy, while the anaphor *themselves* shows semantic agreement, and is further towards the right (i.e. semantic agreement) side of the hierarchy. Conversely, (2c) is less acceptable because the two agreement targets are in the opposite configuration relative to the agreement hierarchy. In other words, the idea is that mismatching agreement is acceptable in the case that the two agreement targets show forms of agreement that coincide with their relative positions on the agreement hierarchy.

However, the relative positions of agreement targets on the agreement hierarchy tend to correlate with both structural and linear distance to the agreement controller. For example, in terms of English number agreement, verbs tend to be closer to their subjects than pronouns or anaphors are to their antecedents (Corbett, 1979; Levin, 2001). Thus, another way to state the generalization is in terms of linear or structural distance:

> We predict that any such constraint will take the form of disallowing a combination of semantic agreement of the nearer element and syntactic agreement of the further. (Corbett, 1979, p. 221)

Based on this observation, we could also formulate an alternative generalization to describe the double mismatch effect, based on the order of agreement targets. According to this generalization, the relative acceptability of (2d) would be due to the fact that the initial element (i.e. the verb) agreeing with the controller shows syntactic agreement, while the subsequent agreeing element (the anaphor) shows semantic agreement, while in (2c), the reverse order of semantic followed by syntactic agreement leads to unacceptability. Note that, according to this idea, it is only the order of singular vs. plural agreement that plays a role in explaining the double mismatch effect, and the parts of speech of the two agreement targets play no role.

These two ways of stating the descriptive generalization suggest two hypotheses regarding the acceptability of double mismatches. According to the *agreement hierarchy-based* account, in the case of a verb and an anaphor agreeing with the same collective noun phrase target, acceptability will be higher if the verb shows singular agreement and the anaphor shows plural agreement, relative to cases where the verb shows plural agreement and the anaphor shows singular agreement. This is because syntactic (singular) agreement of the verb and the semantic (plural) agreement of the anaphor align with the positions of the relevant two dependency types on the agreement hierarchy. In contrast, according to the *order-based* account, where two agreement targets agree with the same collective noun phrase that precedes both targets, acceptability will be higher if the first agreement target shows singular (i.e. syntactic) agreement and the second agreement target shows plural (i.e. semantic) agreement, relative to cases where the first target shows plural agreement and the second shows singular agreement. We will return to discuss the order-based account in more detail in Section 4.

Apart from the *agreement-hierarchy based* account and the *order-based* account, there is a third possible explanation of the pattern of acceptability in (2a–d), based on animacy or humanness, rather than number. According to this explanation, when a committee-type noun participates in plural agreement, it is usually interpreted as denoting a collection of humans, while in singular agreement the canonical interpretation would be as an inanimate entity or institution. As *itself* requires a non-human antecedent, the relative unacceptability of (2c) may be due to a clash in animacy requirements: *avoid*, being plural, biases the interpretation of *committee* towards a group of humans, while *itself* requires *committee* to be inanimate. In contrast, in the relatively acceptable (2d), although *avoids* biases *committee* towards an inanimate interpretation (because it is singular), there is no animacy clash because *themselves* can be used with either an animate or an inanimate antecedent (e.g. *The authors/articles contradicted themselves*). We will return to discuss this *animacy-based* account in 3.6, but for present purposes, it is worth noting that its predictions align completely with those of the *agreement-hieararchy based* account, for the English collective noun examples that we discuss here.

Below, we report two acceptability judgment experiments that are designed to tease apart the *agreement hierarchy-based* and the *order-based* accounts. In Experiment 1, we confirm the intuitive acceptability pattern illustrated in (2a–d). In Experiment 2, we reverse the relative linear order of the agreeing verb and anaphor, in order to compare the two theoretical accounts.

2. Experiment 1

2.1 Participants

A total of eighty native speakers of British English completed the task for payment. Each participant who completed the task received GBP 3.75. Participants were recruited and paid via prolific.co. Of the eighty participants, one could not be included in the analysis because of a failure of data transfer, and seven were removed in order to achieve equal numbers in all Latin Square groups (see below). Participants were removed from over-represented Latin Square groups in reverse chronological order of appearance in the data file, until an equal number of participants per group was achieved.[3] A similar pattern of results was obtained when the full data set was considered. This resulted in a total of seventy-two participants that were included in the analysis. Of these seventy-two, fifty-seven were female, fourteen were male, and one participant preferred not to provide gender information. Mean age was 32 years, with a range of 18–67.

2.2 Stimuli

There were thirty-six items (see 4 for an example set):

(4)
a. **Plural reflexive; Unmarked (past tense) verb:**
The government had distanced themselves from the scandal.

b. **Plural reflexive; Matching (plural) verb:**
The government have distanced themselves from the scandal.

c. **Plural reflexive; Mismatching (singular) verb:**
The government has distanced themselves from the scandal.

d. **Singular reflexive; Unmarked (past tense) verb:**
The government had distanced itself from the scandal.

e. **Singular reflexive; Matching (singular) verb:**
The government has distanced itself from the scandal.

f. **Singular reflexive; Mismatching (plural) verb:**
The government have distanced itself from the scandal.

The design manipulated (i) whether the reflexive was singular (*itself*: (4d–f)) or plural (*themselves*: (4a–c)), and (ii) whether the verb was unmarked for number (*had*, which allows both singular and

[3] The software automatically cycled through latin square groups, by incrementing a variable after each participant had confirmed consent. Unequal numbers in latin square groups can arise if multiple participants start the experiment around the same time. In this scenario, during the time period before the first participant triggers the increment, all participants will be assigned to the same group.

plural subjects: (4a,d)), matched the reflexive in number (e.g. *has … itself*; (4b,e)) or mismatched (e.g. *has … themselves*; (4c,f)). All items used the auxiliary verb *have/had*, indicating pluperfect in the unmarked conditions (*had distanced*) and present perfect in the matching and mismatching conditions (*have distanced*). Note that the design uses verb *matching* (match/mismatch/unmarked) rather than verb *number* (singular/plural/unmarked). The mapping of verb match vs. mismatch onto verb number is summarized in the condition labels of example (4). The subject of each sentence was headed by a collective noun (e.g. *government*). Across the set of 36 items, twenty individual collective nouns were used, with no noun being used in more than two items.

To summarize, the experiment used a 3 × 2 design, manipulating reflexive number (singular vs. plural) and verb matching (match vs. mismatch vs. unmarked). Both of these factors were manipulated within participant and within item.

2.3 Ethics and consent

Ethical approval for this research was granted by the PPLS Research Ethics Committee, University of Edinburgh (257-2021/1).

2.4 Predictions

The *agreement hierarchy-based account* predicts that acceptability would be degraded specifically in the singular reflexive mismatching verb condition. This is because this condition involves a plural verb agreeing semantically with a collective noun, with a singular reflexive agreeing syntactically with the same collective noun, and the low acceptability is due to the fact that the reflexive-antecedent dependency is further towards the semantic end of the hierarchy than the verb-subject dependency—therefore any mismatch in number marking should be less acceptable if it involves plural (semantic) agreement for the verb and syntactic (singular) agreement for the reflexive, relative to the opposite configuration. This pattern should result in an interaction, such that the difference in acceptability between (4f) and (4e) is greater than the difference between (4c) and (4b). In other words, the acceptability penalty for mismatching verbs, relative to matching verbs, should be greater for the singular reflexives than for the plural reflexives. The *order-based account* also predicts an interaction of the same pattern, because the singular reflexive mismatch condition involves an element (the verb) that is linearly closer to the controller participating in semantic (plural) agreement, while the more distant element (the anaphor) participates in syntactic (singular) agreement, and, again, this is predicted to lead to lower acceptability relative to the opposite configuration.

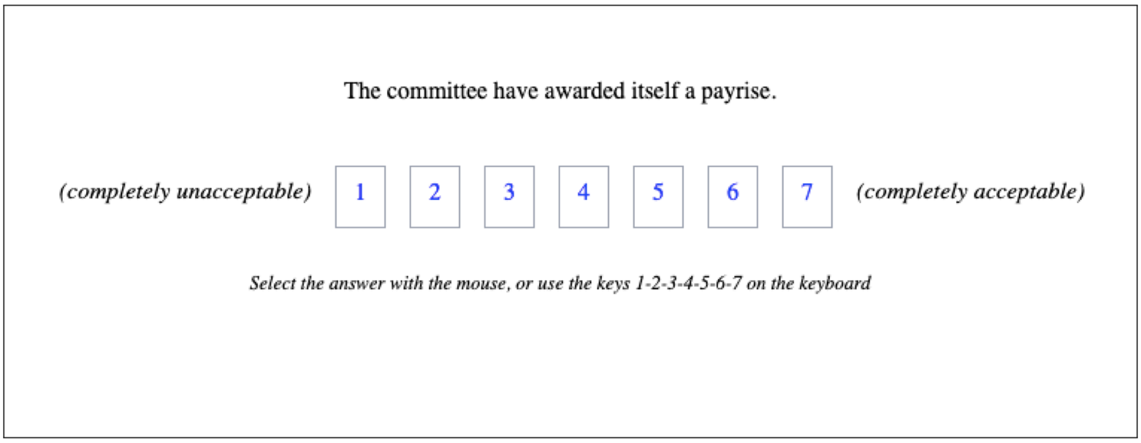
The unmarked verb conditions provide a baseline that controls for alternative explanations of this predicted interaction. One such possibility is that acceptability ratings are degraded in response to a plural verb agreeing with a morphosyntactically singular collective noun (*The* *government have*), relative to a singular verb in the same context (*The government has*). Although such a configuration is described as grammatical in British English, as we have mentioned above, a previous eye-tracking study run on British English speaking participants detected evidence of processing difficulty for plural marked verbs with a morphologically singular committee-type subject (Kreiner et al., 2013, Experiments 2 and 3). Such a tendency may reduce acceptability ratings for (4b) and (4f), thus leading to a confound that could contribute to the predicted interaction, without relying on an explanation that involves specific unacceptability for (4f). Including the unmarked verb conditions allows us to test for such an effect, while controlling for reflexive number. If ratings are affected by a general decrease in acceptability for a plural verb, then, in the plural reflexive conditions, ratings for the verb match condition (4b) should be lower than those for its unmarked baseline (4a), while no such difference should be found in the singular reflexive conditions, between (4d) and (4e). Such a result would manifest itself as an interaction involving these four conditions, and would make the overall results harder to interpret.

2.5 Procedure

The experimental items were combined with 60 fillers, of which half were ungrammatical,[4] and half grammatical. The ungrammatical sentences included both reflexive-antecedent and verb subject number mismatches, as well as other miscellaneous ill-formed sentences. The experiment was implemented on the PCIbex platform (Zehr & Schwartz, 2018), and took place on-line, via a web link from the prolific.co site. The PCIbex software automatically distributed experimental items into a Latin Square, such that each participant was exposed to only one condition from a given item, but across participants, observations were obtained from all conditions of all items, and each participant saw six examples from each condition. On each trial, the participant was asked to judge the acceptability of the given sentence on a 1–7 scale (7 = most acceptable), by clicking the appropriate button with the mouse, or selecting the numerical key on the keyboard (see **Figure 1**). One third of sentences were followed by Yes/No questions, which the participant had to answer by clicking the appropriate answer. After each sentence (or after each sentence-question pair, for stimuli that included questions), a screen was presented for 1000 msec, containing only the string "…" in the top center of the screen.

Following the experiment, the participant was routed back to the prolific.co website, and later received payment. The average time for each participant to complete the task was 14 minutes.

[4] The "ungrammatical" filler sentences included four that used *themselves* with a clearly male or female singular local subject (e.g. *The boy had prepared themselves for the exam.*) These may have been perceived as grammatical by some participants.



The committee have awarded itself a payrise.

Select the answer with the mouse, or use the keys 1-2-3-4-5-6-7 on the keyboard

Figure 1: An example trial.

2.6 Data Analysis

A Bayesian ordinal mixed effects logistic regression was computed on the rating scores using the brms R package (Bürkner, 2017, 2018, 2021), including the two factors of reflexive number and verb matching as fixed effects, and participants and items as random effects. Random slopes were included corresponding to both experimental factors and their interaction, for both participants and items. Random correlation parameters were not included. The contrasts for the fixed effects used dummy coding for the two experimental factors, with the plural reflexive and verb match conditions treated as the respective reference levels. In order to report statistics for all theoretically relevant pairwise comparisons, a second model was run with the reflexive singular and verb match conditions treated as reference levels. The Bayesian models used the brms default priors, including an uninformative flat prior for the *b*-coefficients of all fixed effects. For each model, four chains of 2000 iterations were run, of which the first 1000 iterations were warmup.

2.7 Results

Across all experimental and filler stimuli that included comprehension questions, mean comprehension accuracy was 94% (SE = .88).

Ungrammatical filler sentences received a mean rating of 2.35 (SE = 0.11), and the mean for grammatical filler sentences was 6.60 (SE = 0.05).

Means ratings per condition are displayed in **Figure 2**.

Given that the factor of verb matching had three levels, and the matching verb was the reference level, the model output included (a) comparisons involving the matching and mismatching verb conditions, excluding the unmarked verb conditions (i.e. 4b,c,e,f), and (b) comparisons involving the matching and unmarked verb conditions, excluding the mismatching verb conditions (i.e. 4a,b,d,e). We will call (a) the *main contrast set*, and we will call (b) the *baseline contrast set*. For

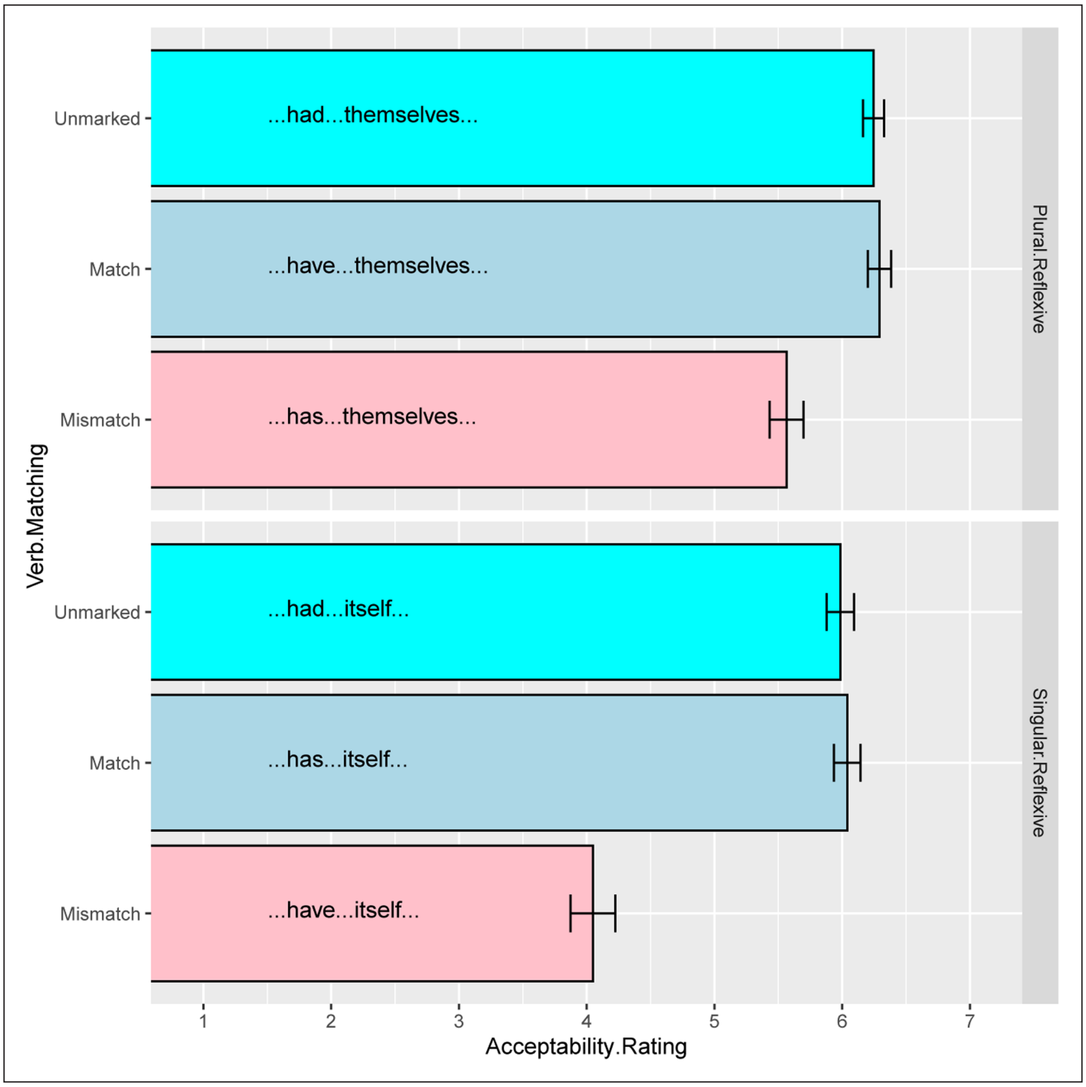


Figure 2: Means ratings (and standard errors) for Experiment 1.

clarity of exposition, we will report the statistical analysis of results separately for the main and baseline contrasts. The main contrast set is used to test for the critical interaction between reflexive number and verb matching, which is predicted by both the *agreement hierarchy-based account* and the *order-based account.*

The *baseline contrast set* is used to test for any differences between the verb match and unmarked verb conditions, and in particular, any differences in the size of such an effect between the singular and plural reflexives (see discussion above). Such a difference in effect sizes would result in an interaction between verb (unmarked vs. match) and reflexive number, and would complicate the overall interpretation of the results.

Main contrast set: Bayesian model results for the *main contrast set* are summarized in **Table 2**:

Table 2: MAIN CONTRAST SET: Bayesian model estimates for effects involving **Verb Match and Mismatch** conditions (i.e. excluding unmarked verb conditions). Output shows relevant paired comparisons (X vs. Y) and interaction (X × Y).

Effect	**Estimate**	**95% CrI**	***P*(*b* < 0)**
Mismatching vs. Matching verb (at plural reflexive)	–1.10	[–1.47, –0.72]	.999
Mismatching vs. Matching verb (at sing. reflexive)	–2.53	[–2.96, –2.09]	.999
Verb matching × Refl. number	–1.50	[–2.10, –0.91]	.999

The most important result from the *main contrast set* is the interaction of verb matching by reflexive number. There was strong evidence for this interaction (see **Table 2**). This interaction showed that the rating penalty for the mismatching verb relative to the matching verb was greater for the singular reflexives (4f) vs. (4e) than for the plural reflexives (4c) vs. (4b). This therefore provides strong support for the prediction of both the *agreement hierarchy-based account* and the *order account* that acceptability would be specifically reduced for the mismatching verb with the singular reflexive. Despite the difference in the size of this verb-mismatch cost, pairwise comparisons showed strong evidence for the mismatch cost at both levels of the reflexive number factor (see **Table 2**).

Baseline contrast set: Bayesian model results for the baseline contrast set are reported in **Table 3**:

Table 3: BASELINE CONTRAST SET: Bayesian model estimates for effects involving **Verb Match and Unmarked** conditions (i.e. excluding mismatching verb conditions). Output shows relevant paired comparisons (X vs. Y) and interaction (X × Y).

Effect	**Estimate**	**95% CrI**	***P*(*b* < 0)**
Matching vs. Unmarked verb (at plural reflexive)	–0.08	[–0.42,0.26]	.68
Matching vs. Unmarked verb (at sing. reflexive)	–0.04	[–0.36,0.29]	.61
Verb (unmarked vs. match) × Refl. number	0.06	[–0.36, 0.50]	.39

In the baseline contrast set, we use the unmarked verbs as a baseline to test for the cost of a plural marked verb, relative to a singular marked verb in the same context. If a singular verb reduces acceptability, then in the plural reflexive conditions, the matching verb condition (4b) should have lower ratings than the unmarked verb condition (4a), while no such difference should be observed between the matching and unmarked verbs in the singular reflexive conditions (4d) and (4e), and this should lead to an interaction between verb (unmarked vs. match) and reflexive number. As seen in **Table 3**, there was no clear evidence of such an interaction, nor was there any evidence for a difference between the verb match and unmarked verb condition at either of the two levels of the reflexive number factor. Thus, the main interaction that we report in **Table 2** above cannot be explained by an acceptability cost for a verb with singular morphology.

2.8 Discussion

The results confirmed the pattern described in Section 1 as the *double mismatch effect*. That is to say, in cases where the verb and reflexive mismatched in agreement with the collective noun agreement controller, acceptability was higher when the verb showed singular agreement and the reflexive showed plural agreement than in the reverse configuration. This result is compatible with an explanation based on either the agreement hierarchy-based account or the order-based account. In Experiment 2, we reversed the relative linear order of the anaphor and the verb with respect to Experiment 1, so that the predictions of the two accounts diverged.

3. Experiment 2

3.1 Participants

A total of 92 native speakers of British English completed the task for payment. Each participant who completed the task was paid GBP 3.75. None of the participants had participated in Experiment 1. Participants were recruited and paid via prolific.co. Twenty participants were removed from the analysis in order to achieve balanced Latin Square groups. As in Experiment 1, participants were removed in reverse chronological order of appearance in the PCIbex data file, until an equal number of participants per group was achieved. The analysis therefore included seventy-two participants. A similar pattern of results was obtained when the full data set was considered. Of these seventy-two, fifty-six were female and sixteen were male. Mean age was 39 years, with a range of 20–74.[5]

3.2 Stimuli

Forty stimuli were included in the experiment (see (5) for an example set). The design manipulated the number of the reflexive (singular vs. plural), and whether or not the verb matched that number (verb-match vs. verb-mismatch). The stimuli were designed so that the reflexive appeared before the verb, thus reversing the relative order of these two elements in comparison with Experiment 1. In order to achieve this ordering, we required more complex sentence structures than those that were used in Experiment 1. We used two sentence structures. The first used a

[5] Age information was unavailable for one participant.

relative clause modifying the main clause subject (5a–d), with the reflexive appearing inside this relative clause, and the agreeing verb appearing in the main clause. The verb inside the relative clause used past tense, thus avoiding number morphology. The second sentence construction used coordination (5e–h), with the reflexive appearing in the first conjunct and the agreeing verb appearing in the second. We assumed that coordination was at a level of structure that excluded the matrix subject, which was shared between conjuncts, though see below for discussion of alternative analyses. In the coordination construction, the verb of the first conjunct used past tense, and did not indicate number agreement (e.g. *distanced*), while the second conjunct used the tensed auxiliary verb *was/were*, showing number agreement. The sentence type (relative clause vs. coordination) was included as a within-item factor in the design, in order to allow conclusions that generalize across these two different ways of allowing the reflexive-verb order.

(5) a. **Relative Clause; Plural reflexive; Matching (plural) verb:**
The government that distanced themselves from the scandal were discussed on the news.

b. **Relative Clause; Plural reflexive; Mismatching (singular) verb:**
The government that distanced themselves from the scandal was discussed on the news.

c. **Relative Clause; Singular reflexive; Matching (singular) verb:**
The government that distanced itself from the scandal was discussed on the news.

d. **Relative Clause; Singular reflexive; Mismatching (plural) verb:**
The government that distanced itself from the scandal were discussed on the news.

e. **Coordination; Plural reflexive; Matching (plural) verb:**
The government distanced themselves from the scandal and were discussed on the news.

f. **Coordination; Plural reflexive; Mismatching (singular) verb:**
The government distanced themselves from the scandal and was discussed on the news.

g. **Coordination; Singular reflexive; Matching (singular) verb:**
The government distanced itself from the scandal and was discussed on the news.

h. **Coordination; Singular reflexive; Mismatching (plural) verb:**
The government distanced itself from the scandal and were discussed on the news.

The design was therefore a 2 × 2 × 2, with factors sentence type (relative clause vs. coordination), reflexive number (singular vs. plural) and verb matching (verb-match vs. verb-mismatch), and all three factors were manipulated within item and within participant. The critical hypothesis is tested by the interaction of reflexive number by verb matching. The *order-based account* predicts such an interaction: the cost for a mismatching verb (relative to a matching one) should be greater in the plural reflexive conditions (where initial semantic agreement between the plural reflexive and collective subject is followed by syntactic agreement between the singular verb and the subject), than in the singular reflexive conditions (where initial morpho-syntactic agreement is followed by semantic agreement). The *agreement hierarchy-based account* also predicts an interaction between reflexive number and verb matching, but with a reverse pattern. The verb-mismatch cost should be greater with the singular reflexives than with the plural reflexives, because in the singular reflexive case, the mismatching verb establishes semantic agreement with the collective noun phrase, while the reflexive, which is further towards the semantic end of the agreement hierarchy, participates in syntactic agreement. In contrast, plural reflexives should lead to a smaller cost for the verb mismatch, because in this case, the verb participates in syntactic agreement and the reflexive in semantic agreement, and this corresponds to the ordering of these two dependency types on the agreement hierarchy. Given the contrast coding used in the analysis, the *order-based account* predicts the estimate for the interaction of reflexive number and verb matching to have a positive sign, and the *agreement hierarchy-based account* predicts the interaction to have a negative sign.

One caveat that has to be stated about this experiment is that both relative clause and coordination versions of the stimuli could be argued to include empty elements that may also bear a value for a number feature. For the coordination stimuli, although we assumed the coordination of a constituent below the level of the sentence, with a shared subject, there are also analyses that would assume we are dealing with coordination at the sentence level (Van Valin, 1986, 3), with an empty pronominal element as the subject of the second conjunct (though see Godard (1989) for arguments against such an analysis):

(6) $[_S$ [The government]$_i$ awarded itself a payrise] and $[_S$ *pro*$_i$ were ….]

If this analysis is correct, it would imply that the dependency between the auxiliary verb *were* and the collective noun phrase is indirect, and is mediated by the empty pronominal (i.e. there is a dependency between *were* and pro, and another between pro and *the government*).

The relative clause may also include an empty element, namely a relative pronoun operator (though see Kayne (1994), Vergnaud (1974) for accounts that do not involve this element):

(7) [The government]$_i$ [RELPRO$_i$ that t$_i$ awarded itself a payrise] were …

If this analysis is correct, it would also imply that the dependency between the reflexive and the collective noun phrase is indirect, being mediated in this case by the empty relative pronoun operator.

In designing Experiment 2, we assumed that the crucial elements for deriving the predictions of the agreement hierarchy-based account were the overtly marked auxiliary *were/was* and reflexive *themselves/itself*, and that these predictions would be unaffected by the existence of any (putative) empty categories within the sentence structure. However, we acknowledge that this assumption may not hold, and we return to the issue below in 3.6, where we will consider two further hypotheses that assign a more important role to the empty categories. We will argue that neither of these hypotheses can explain the overall pattern of results.

3.3 Procedure

The procedure was identical to that of Experiment 1 in all relevant respects. The Latin Square procedure resulted in any given participant being presented with five items from each of the eight conditions. The 60 fillers used in Experiment 1 were lengthened, given the longer experimental stimuli in Experiment 2 than Experiment 1. Other characteristics of the fillers (e.g. regarding grammaticality) remained the same. The average time for each participant to complete the task was 19 minutes.

3.4 Data Analysis

A Bayesian ordinal mixed effects logistic regression was computed on the rating scores. As in Experiment 1, the brms R package was used (Bürkner, 2017, 2018, 2021), including the three factors of reflexive number, verb matching and structure as fixed effects, and participants and items as random effects. Random slopes were included corresponding to all three experimental factors and their interactions, for both participants and items. Random correlation parameters were not included. The experimental factors were coded using sum coding, with the two levels of each factor coded as –0.5 and 0.5 respectively.

3.5 Results

Across all experimental and filler stimuli that included comprehension questions, mean comprehension accuracy was 92% ($SE = .69$).

Ungrammatical filler sentences received a mean rating of 2.55 ($SE = 0.12$), and the mean for grammatical filler sentences was 6.54 ($SE = 0.05$).

Mean ratings per condition are displayed in **Figure 3**.

Bayesian model output results are shown in **Table 4**:

From the point of view of testing the theoretical predictions, the salient result was the interaction of verb matching by reflexive number, for which there was strong evidence. This interaction had a positive sign, meaning that the acceptability penalty for mismatching (vs. matching) verbs was greater for the plural reflexive than for the singular reflexive, consistent with the *order-based account*, and not with the *agreement hierarchy-based account*. We will use further contrast analyses to explore this interaction in more detail below. The analysis also showed evidence for a main effect of reflexive number (with overall higher ratings for singular reflexives than for plural reflexives), a main effect of structure (coordination structures had overall higher ratings than relative clause structures), and a main effect of verb match (matching verbs had

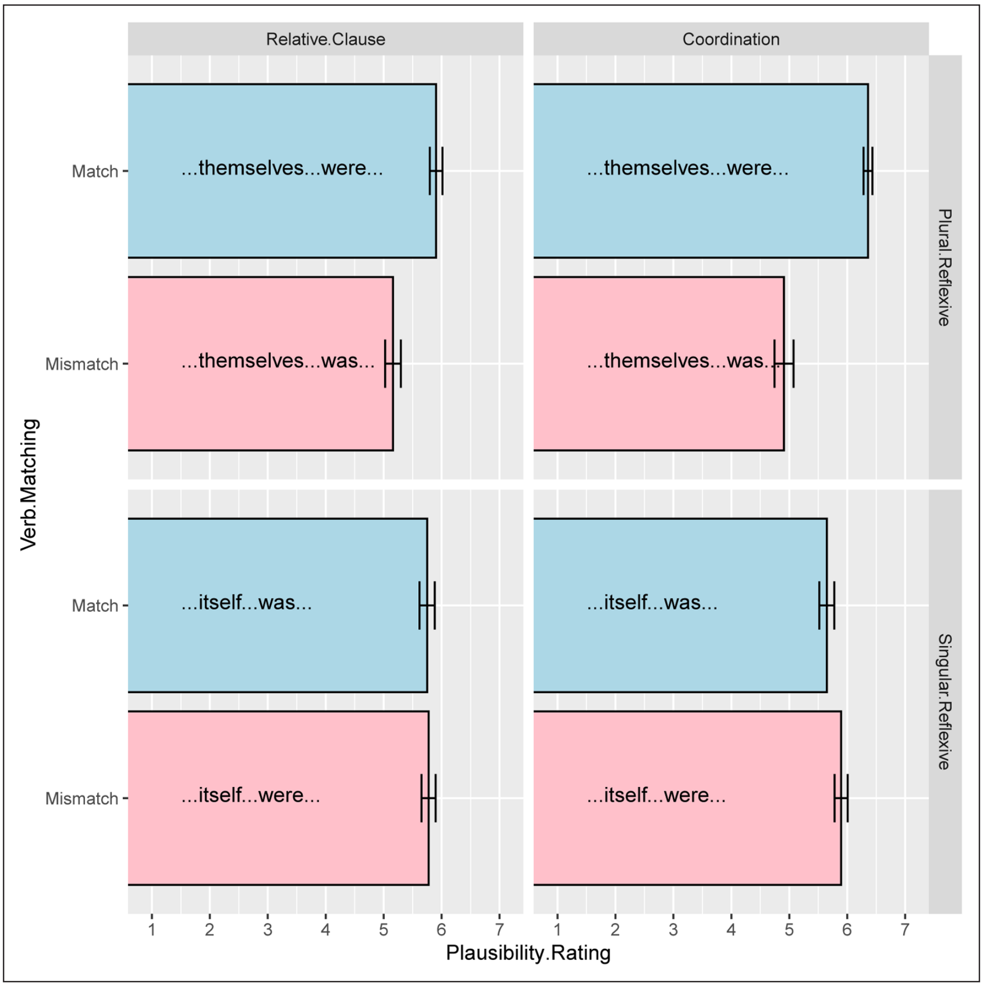


Figure 3: Means ratings (and standard errors) for Experiment 2.

Table 4: Bayesian model output for Experiment 2, showing **(a)** estimates for main effects and interactions, **(b)** credible intervals, and **(c)** the probability that the effect is greater than zero (for positive coefficients), or less than zero (for negative coefficients).

Effect	**Estimate**	**95% Credible interval**	$P(b>/<0)$
Reflexive number	0.24	[0.03, 0.45]	.99
Structure	–0.19	[–0.37, –0.02]	>.999
Verb matching	–0.79	[–1.04, –0.54]	>.999
Verb × Reflexive	2.06	[1.47, 2.67]	>.999
Verb × Structure	0.42	[0.09, 0.75]	.99
Reflexive × Structure	0.16	[–0.20, 0.50]	.8
Verb × Reflexive × Structure	–1.60	[–2.42, –0.80]	>.999

higher ratings than mismatching verbs). As well as these effects, there was evidence for a two-way interaction of verb-matching by structure, indicating that the penalty for a mismatching verb was greater for the coordination structure than for the relative clause structure. There was no clear evidence for a two-way interaction of reflexive number by structure. Finally, there was evidence for a three-way interaction among the experimental factors, which can be interpreted as showing that the effect size of the two-way interaction of verb-matching by reflexive number was greater for the coordination structure than for the relative clause structure (see **Figure 3**).

The evidence for the three-way interaction justified separate analyses for the coordination and relative clause structures, as it was important to establish that the predicted interaction of verb-matching by reflexive number could be generalized across sentence types, even if the size of this effect may differ between them. The separate analyses indeed showed clear evidence of the two-way interaction for both sentence types (Coordination: $b = 2.94$, CrI = [2.16,3.78], $P(b>0) > .999$; Relative clause: $b = 1.27$, CrI = [0.54,1.98], $P(b>0) > .999$). These interactions were then further analyzed using contrast models that tested the verb-matching effect at each level of the reflexive number variable, within each sentence type. For both sentence types, there was a clear degradation in acceptability for mismatching verbs (relative to matching verbs) when the reflexive was plural (Coordination: $b = 2.51$, CrI = [1.97,3.06], $P(b>0) > .999$; Relative clause: $b = 1.19$, CrI = [0.81,1.57], $P(b>0) > .999$). However, there was no such evidence for such a verb-mismatch cost at the singular reflexive level for the relative clause structure, and for the coordination structure, if anything there was weak evidence for a cost for the *matching* verb (Coordination: $b = –0.44$, CrI = [–0.91,0.05], $P(b<0) = .97$; Relative Clause: $b = –0.05$, CrI = [–0.51,0.44], $P(b<0) = .58$).

3.6 Discussion

The central result of Experiment 2 was the two-way interaction of verb-matching by reflexive number. The pattern of this interaction was such that there was a verb-mismatch cost when the reflexive was plural, but the lack of such a cost when the reflexive was singular. There was clear evidence for this interaction for both sentence structures examined here, even though the magnitude of the effect was greater for the coordination sentences than for the relative clauses. The pattern is not consistent with the agreement hierarchy-based account, which would have predicted that the mismatch cost would have been greater for the singular reflexive condition than for the plural reflexive condition. Instead, the results are consistent with the order hypothesis, given that the first agreeing element was the reflexive. When this reflexive was plural, this established an initial semantic agreement, leading to degraded perceived acceptability when a mismatching (i.e. singular) verb later signalled morpho-syntactic agreement. In contrast, when the reflexive was singular, morpho-syntactic agreement was initially established, and a mismatching (i.e. plural) verb could subsequently signal semantic agreement without a large degradation in acceptability. It is also worth noting that the results are inconsistent with the *animacy-based* account that we briefly discussed in Section 1. According to that account, there should have been an acceptability cost for conditions in which a plural verb *were* is combined with a singular reflexive *itself*, due to the clashing animacy/humanness requirements. However, this was not found, and the lower acceptability was found instead for the case in which there was a combination of a singular verb *was* and a plural reflexive *themselves*. The results also showed that this two way interaction effect was larger for the coordination sentence structure than it was for the relative clause structure. This effect was not predicted, and any explanation would need to be speculative. We will therefore not discuss it further.

One potential concern with Experiment 2 is that a large number of the experimental stimuli had a noun phrase intervening between the critical reflexive and the number-marked auxiliary verb (e.g. *The government that distanced themselves from* ***the scandal*** *were…*). This was the case for 35 of the 40 experimental items, and since the phrase was inevitably either singular (32 items) or plural (3 items), there was a potential for number attraction or interference effects to influence the results. To investigate this possibility, we isolated the five items that did not include an intervening noun phrase in this position (these items used adverbials: (e.g. *The crowd that conducted themselves* ***very well*** *were…*). Taking the acceptability ratings for these five items, we ran a new Bayesian model. This model used a simplified fixed and random effects structure, to allow for the much reduced dataset— we used only reflexive number and verb matching and their interaction for the fixed effects, and only participant-related random effects. The results of this analysis replicated the crucial interaction between reflexive number and verb matching that was found in the main analysis ($b = 1.84$, $CrI = [0.78,2.95]$, $P(b>0) > .999$), with very similar means to those of the main analysis (Plural-reflexive/Verb-match: (Mean = 6.20, SE = 0.12); Plural-reflexive/Verb-mismatch: Mean = 5.10, SE = 0.22; Singular-reflexive/Verb-match: Mean = 5.72, SE = 0.15; Singular-reflexive/Verb-mismatch: Mean = 5.74, SE = 0.18). Thus, even if the intervening NP led to interference effects, this is unlikely to have explained the main result.

A second potential concern with Experiment 2 is related to the possibility that either the relative clause or the coordination sentences may have included empty elements, as mentioned above. We assumed that, even if present, these elements would not play an important role in the acceptability of agreement mismatches. In what follows, we discuss two possible hypotheses that do not make this simplifying assumption.

The first hypothesis is what we will call the *independent domains* hypothesis. According to this hypothesis, in the relative clause conditions, agreement within the relative clause is controlled independently by the empty relative pronoun (see (7) above), while in the coordination conditions, agreement within the second conjunct is controlled independently by the empty pronominal subject (see (6) above). Thus, in each case, number agreement within the domain that includes the empty element (relative clause or second conjunct) may be independent of number agreement in the local domain of the committee-type noun phrase (the main clause in the relative clause conditions, or the first conjunct in the coordination conditions). If the number agreement processes are truly independent,[6] then any preference for the number marking of the verb (singular vs. plural) should be independent of the number marking of the reflexive, and vice versa. Assuming that independence maps onto additivity in acceptability ratings, then this would be falsified by a main effect of verb matching. This is because any mismatch effect for the reflexive singular conditions should be cancelled out by an equal and opposite mismatch effect for the reflexive plural conditions: for example, a preference for matching (vs. mismatching) verbs within the reflexive singular conditions would indicate a preference for singular marked verbs over plural marked verbs. An equivalent preference for singular verbs within the reflexive plural conditions would lead to a preference for *mismatching* (vs. matching) verbs. In a situation where preferences for reflexive and verb number marking are truly independent, these two opposing preferences would be of equal size, and would cancel out the main effect of verb matching.[7] However, as seen in **Table 4**, there was in fact very strong evidence for the main effect of matching. Of course, it is possible that only one of the relative clause or coordination sentence types includes empty categories, meaning that the independent domains hypothesis might apply to only one of them. However, there was strong evidence for the main effect of matching in both of the smaller models that were run on relative clause and coordination conditions separately (Relative clause: b = –0.54, CrI = [–0.79,–0.29], $P(b<0)$ > .999; Coordination: b = –1.03, CrI = [–1.41,–0.67], $P(b<0)$ > .999). Thus, the results for Experiment 2 do not support the *independent domains* hypothesis.

The second hypothesis that we consider is what we call the *empty element agreement hierarchy* hypothesis. This hypothesis is the same as the *agreement hierarchy-based* hypothesis, except that the important elements with respect to the agreement hierarchy are the empty categories, rather than the number-marked auxiliary or reflexive. This would mean, for the coordination conditions, that the double mismatch effect involves the interplay between the reflexive in the first conjunct and the empty pronominal in the second conjunct (and the number marking of *was/were* plays no role beyond matching the number feature of the empty pronominal subject). For the relative clause conditions, it would mean that the double mismatch effect involves the interplay between

[6] Assuming full independence is almost certainly an oversimplification. However, in examples like *It is me who has kept out of trouble*, the embedded verb *has* does not match the person features of the head *me* (Akmajian, 1970; Furuya, 2017), suggesting that (empty) relative pronouns in English may in some circumstances be at least partially independent from their heads in terms of phi-features.

[7] Note that this prediction is conceptually easier to grasp if we assume an alternative labelling of conditions where the factors are reflexive number (singular vs. plural) and verb number (also singular vs. plural, rather than match vs. mismatch). In this case, the hypothesis would be falsified by an interaction between reflexive and verb number. This interaction in the alternative labelling of conditions is in fact mathematically equivalent to the main effect of matching in the labelling of conditions that we have adopted in this paper.

the empty relative pronoun in the relative clause and the auxiliary in the main clause (and the number marking on the reflexive in the relative clause plays no role beyond matching the number feature of the empty relative pronoun). If this were the case, and assuming that the agreement hierarchy can be applied to empty elements, then for the coordination conditions, the *empty element agreement hierarchy hypothesis* makes no specific prediction for the double mismatch effect. This is because both the empty pronominal in the second conjunct and the reflexive pronoun in the first conjunct would occupy the "personal pronoun" position of the agreement hierarchy (see **Table 1**), and thus would not differ in their propensity for semantic vs. syntactic agreement. This means that there would be no prediction for either one of the two mismatch configurations to be more acceptable than the other. However, this would not explain our results, as we did indeed find an asymmetry. For the relative clause stimuli, the predictions of the *empty element agreement hierarchy account* are unchanged from those of the *agreement-hieararhcy account* given our original assumptions. This is because in the agreement hierarchy, relative pronouns occupy a position that is further towards the semantic end of the spectrum than the position occupied by verbs (see **Table 1**). This means that greater acceptability will be expected where the mismatch mirrors the relative positions of the elements on the agreement hierarchy than where it does not; in other words, sentences will be judged to be more acceptable if the relative pronoun participates in semantic (plural) agreement and the auxiliary participates in syntactic (singular) agreement, relative to the opposite configuration. This prediction is the same as the one outlined above for the *agreement hierarchy-based* account, where we assumed that the relevant elements were the reflexive and the auxiliary, because the reflexive is again further towards the semantic side of the hierarchy than the verb. The *empty element agreement hierarchy* hypothesis is therefore inconsistent with the relative clause results for Experiment 2—as relative pronouns occupy a more semantically oriented slot than verb subject agreement, the prediction would be that acceptability would be higher where the verb-subject dependency shows singular agreement while the relative pronoun-head noun dependency (indicated by overt agreement on the reflexive) shows plural agreement, relative to the opposite configuration. However, this is the reverse of the results that were obtained.

To summarize, although the interpretation of Experiment 2 is inevitably complicated by the possible involvement of empty categories, we argue that the results are not compatible with either the *independent domains* hypothesis or the *empty element agreement hierarchy* hypothesis. Instead, the *order-based* hypothesis provides the best account of the results.

Of course, it may be possible to devise alternative hypotheses, involving empty agreeing elements. However, in order to account for the results of Experiment 2, any such explanation would need to account for the fact that the qualitative pattern of results is the same for both the coordination and relative clause conditions, even though the position of the putative empty element is different for each structure, and would presumably affect the first (reflexive) agreement computation for the relative clauses but the second (verb) agreement computation for the coordination sentences. Ultimately, the findings reported here should be replicated using a wider variety of sentence types. One possibility would be to use VP topicalization in English to allow for a reflexive to appear before the agreeing auxiliary verb (e.g. *Criticize itself/themselves though the government never has/have, there have been plenty of grounds for self-criticism.*[8]). This differs from the examples used here in Experiment 2 in that it arguably avoids the complication of having empty elements that might be assigned features independently. Note also that another difference from the stimuli used in Experiment 2 is that in the VP-topicalization example, the reflexive precedes not only the inflected verb but also the collective noun phrase. Another possibility would be to look at other forms of semantic agreement, and in other languages, preferably including those that allow the order of agreeing elements to be manipulated more flexibly without introducing large differences in underlying sentence structure.

4. General Discussion

Experiment 1 provided a confirmation of the double mismatch effect using controlled experimental methods, in cases where the verb precedes the anaphor. This result was compatible with either the agreement hierarchy-based account or the order-based account. Experiment 2 reversed the linear order of verb and reflexive, and showed a pattern that is more compatible with the order-based account.

Before we go on to discuss the theoretical implications, there is one aspect of the results that deserves comment, which is the fact that the overall acceptability penalty for a mismatch between verbs and anaphors was greater in Experiment 1 than it was in Experiment 2. This can be seen from the fact that in Experiment 1, both singular and plural reflexives showed strong evidence for an acceptability penalty when the number mismatched that of their respective verbs, even though the mismatch penalty was larger for singular reflexives than for plural reflexives. In contrast, in Experiment 2, there was a detectable mismatch penalty only for the plural reflexives, and no evidence of a penalty for the singular reflexives. Moreover, for the dispreferred mismatching condition (singular reflexives in Experiment 1 and plural reflexives in Experiment 2), the numerical size of the mismatch effect was larger in Experiment 1 than it was in Experiment 2 (compare **Figures 2** and **3**). Explanations for this difference must be speculative in the absence of further empirical studies. One possibility is that, due to the more complex sentences of Experiment 2, people engaged in more shallow processing in that experiment relative to Experiment 1, and were thus less sensitive to mismatches overall. This shallow processing explanation would be supported by a lower rate of comprehension accuracy in Experiment 2 relative to Experiment 1. However, although comprehension accuracy was

[8] I am grateful to an anonymous reviewer for suggesting this idea.

numerically slightly lower in Experiment 2, the difference was small, and rates were relatively high in both experiments. This was the case even when considering only experimental stimuli (i.e. ignoring fillers), where the mean accuracy rate was 94% (*SE* = 0.89%) for Experiment 1 and 91% (*SE* = 0.88) for Experiment 2. Thus, even if the participants of Experiment 2 employed shallower processing than those of Experiment 1, this was not to such a degree as to have a large effect on comprehension accuracy.

It is possible that the difference in the size of the mismatch effect in Experiment 2 relative to Experiment 1 is related to the difference in structural distance between the verb and the reflexive in Experiment 1 relative to Experiment 2. In Experiment 1, the anaphor appeared in the verb phrase local to the number-marked auxiliary verb. In Experiment 2, however, the anaphor appeared in a separate verb phrase from the one that was local to the number-marked auxiliary. It may be the case that number mismatches between two agreement targets sharing a controller are less acceptable when the two targets are local to each other than when they are not.

We now move on to consider the interpretation of the main results of this paper. The first question is whether the pattern of results that we report could be captured at the competence level. A competence-based explanation would mean that certain types of agreement mismatch could be generated by the competence grammar while other types could not, and that this distinction would be reflected in acceptability ratings. In order to consider this question, we will discuss some of the proposals that have been made about collective noun agreement in the theoretical linguistics literature. We will see that the theories can be divided into those that allow agreement mismatches of the type that we examine in this paper and those that do not. Of those that allow agreement mismatches, none of the proposals that we discuss can account for the full range of acceptability contrasts reported in this paper between different types of mismatches. We then move on to sketch a tentative processing-based explanation based on incremental comprehension.

In Head-Driven Phrase Structure Grammar, henceforth HPSG (Pollard & Sag, 1994), agreement features, including those for number, are attributes of referential *indices*, described by the authors as “abstract objects that function in discourse to keep track of the entities that are being talked about” (Pollard & Sag, 1994, p. 67). The object denoted by a collective noun like *government* may be assigned either a singular or a plural index, depending on whether the referent is interpreted as a non-aggregate entity, or as an aggregate of multiple entities, and this leads to singular or plural agreement respectively. If this aspect of the interpretation changes, a referent may be assigned a new index, so that, for example, the index assigned to the entity denoted by *government* may change from singular to plural. This can account for cases where the number agreement changes from one sentence to the next (e.g. *The government was unpopular. They failed to produce good policies.*) However, Pollard and Sag (1994) predict that the types of within-sentence double mismatch that we consider in this paper (e.g. (4c,f)) should be ungrammatical. Consider these examples repeated below:

(4) c. The government has distanced themselves from the scandal.
f. The government have distanced itself from the scandal.

In HPSG (Pollard & Sag, 1994), an anaphor like *themselves* in (4c), or *itself* in (4f) is co-indexed with the antecedent *the government*, via structure sharing. As indices bear number features, this index should therefore be plural in (4c) and singular in (4f). However, the process of subject-verb agreement requires that the same index be singular in (4c) and plural in (4f), resulting in a clash of features and thus ungrammaticality in both cases. Thus, HPSG does not allow either type of mismatch considered here to be generated by the grammar. The proposal of Pollard and Sag (1994) therefore does not capture the critical constrasts that we report in this paper, simply because no type of mismatch is generated by the competence grammar.

The second account that we consider is a proposal by den Dikken(2001). According to this proposal, a phrase like *the government* may be headed by a null plural pronominal element. When this pronominal element is present, it participates in dependencies with agreement targets, such as verbs or anaphors, resulting in plural agreement. In contrast, when the pronominal element is not present, agreement takes place directly with the (morphologically singular) collective phrase, and singular agreement is the result. Thus, if there are two agreement targets that form dependencies with a single collective controller within the same sentence, then agreement may either be singular (pronominal absent) or plural (pronominal present), but not both. Thus, den Dikken (2001) predicts both mismatching examples (4c) and (4f) to be ungrammatical, so again, this proposal does not allow for the acceptability contrast between these two to be captured at the competence level.

We now consider theoretical approaches which allow for double mismatches to be generated within the competence grammar. Sauerland and Elbourne (2002) achieve this by assuming that number agreement in English involves two different categories of features, either of which can take a singular or a plural value. The *number* feature reflects whether there is one entity or more than one entity under discussion, and the singular or plural value for this feature will correlate with the morpho-syntactic marking on the noun. For example, *government* corresponds to a single entity and *governments* to multiple such entities. The *mereology* feature reflects whether or not the entity is being interpreted as consisting of more than one member, and the value of this feature is not reflected in the morphology of the noun. For example, *government* may take either a singular or a plural value for the mereology feature, depending on its interpretation. Double mismatches are then handled by allowing, for example, verb-subject agreement to occur via the number feature and anaphor-antecedent agreement to occur via the mereology feature, or vice versa. Thus, Sauerland and Elbourne's approach allows both (4c) and (4f) to be generated by the competence grammar. Therefore, the proposal does not capture the acceptability difference between (4c) and (4f) at the competence level.

The final theoretical account that we consider is the proposal by P. Smith (2015, 2017). According to Smith's proposal, a number feature may be split into two halves, of which one (which Smith calls iF) is interpreted by semantics, and the other (called uF) is interpreted by morphology. A collective noun like *government* may include a singular value for uF and a plural value for iF. Double mismatches are then handled by different agreement targets interacting with different parts of the feature, bearing different values, for example, a verb agreeing with the uF and an anaphor with iF. Unlike the theoretical accounts mentioned above, Smith's proposal does predict a difference in acceptability between (4c) and (4f), with (4c) being predicted to be more acceptable, as confirmed by the results of Experiment 1 above. Smith's explanation of this asymmetry is based on the principle of *valuation economy*, which requires that, if two targets agree with the same controller in the same domain, then either both need to agree via uF or both need to agree via iF. The double mismatch in (4c) is the result of agreement occuring in two different domains, allowing uF to be used in one domain and iF in another, without violating valuation economy. Specifically, the (singular) verb-subject agreement occurs via uF post-syntactically (in the branch to the P(honetic) F(orm) component of grammar, in Smith's derivational Minimalist-based system), and the (plural) anaphor-antecedent agreement occurs via iF, within the syntactic component. We now explain why the mismatch in (4f) is ruled out in this system. First, it is assumed that anaphor-antecedent agreement has to occur within the syntactic component, to allow the dependency to be interpreted semantically, while verb-subject agreement may occur either within the syntactic component or post-syntactically (i.e. in the PF branch of the grammar). Given that the anaphor-antecedent agreement in (4f) is singular, it must have been derived via uF, and, being an anaphor-antecedent dependency, it must have occurred within the syntactic component of the grammar. However, for the verb-subject agreement to be plural, it must have occurred via iF, but in order to satisfy valuation economy, this agreement cannot take place within the syntactic component, because a uF agreement has already taken place there, involving the same controller. But there is no possibility for this agreement to take place in the PF branch of the grammar, because, according to the theory iFs are visible only to agreement within the syntactic component.

Given the discussion above, it is clear how the proposal of P. Smith (2015, 2017) can be used to explain the acceptability asymmetry between between (4c) and (4f) that we found in Experiment 1. However, the proposal makes the incorrect prediction for Experiment 2, where the linear order of the agreeing verb and anaphor is reversed. This is because the relevant distinctions are based on the types of agreement targets (i.e. verb vs. anaphor), rather than on linear order. In fact P. Smith (2015) discusses two examples that are similar in all relevant respects to our mismatch conditions in Experiment 2, with the following predicted grammaticality judgements (P. Smith, 2015, p. 232):[9]

[9] These predicted judgements from P. Smith (2015) are based on a refinement of the valuation economy proposal based on derivation order, but the predictions are the same as for the version that we described above.

(8) a. The committee that gave themselves a hefty payrise is being indicted on charges of corruption.

b. *The committee that gave itself a hefty payrise are being indicted on charges of corruption.

In P. Smith (2015), (8b) is predicted to have lower acceptability than (8a), for analogous reasons that (4f) is predicted to be less acceptable than (4c). However, the results of Experiment 2 show that in fact, sentences like (8b) are judged to be *more* acceptable than those like (8a).

To summarize, of the four theoretical proposals that we have discussed so far, none can capture our results in terms of grammaticality at the competence level. How, then, might the results be explained? One way to think about this is to consider how morpho-syntactic and semantic agreement processes might unfold in real time during sentence comprehension. In a study on German collective nouns, Schweppe (2013) showed that people's preferences for pronoun-antecedent agreement tended to shift from syntactic to semantic agreement as the distance from the agreement controller increased. For example, in her Experiment 2 (self-paced reading), Schweppe (2013) examined sentences like the following:

(9) a. **Singular/Plural pronoun; short distance**
Das Militär war noch immer preußisch organisiert. Es/Sie legte Wert auf eine kaisertreue Gesinnung. Daran hatte sich nichts geändert.
'The army was still organized in a Prussian way. It set a high value on loyalty to the emperor. Nothing had changed.'

b. **Singular/Plural pronoun; long distance**
Das Militär war noch immer preußisch organisiert. Daran hatte sich nichts geändert. Es/Sie legte Wert auf eine kaisertreue Gesinnung.
'The army was still organized in a Prussian way. Nothing had changed. It set a high value on loyalty to the emperor.'

In the design, Schweppe (2013) manipulated whether the pronoun referring to a collective noun was singular (*Es*) or plural (*Sie*). She also manipulated whether or not an intervening clause (*Daran hatte sich nichts geändert* 'Nothing had changed') intervened between the sentence containing the pronoun and the sentence containing the collective noun, resulting in long distance (intervening sentence) and short distance (no intervening sentence) conditions. On the word following the verb in the critical pronoun sentence, pronoun number and distance interacted, such that, for the singular pronoun, reading times were longer in the long distance than the short distance condition, whereas no such difference was found for the plural pronouns (with a numerical trend in the opposite direction). In a production experiment in the same paper, using a "fill in the gap" task, Schweppe (2013) also found that the proportion of singular pronouns (relative to plurals) reduced in the long distance condition relative to the short distance condition. These findings are compatible with corpus results showing that rates of plural agreement for pronouns referring to collective nouns in English increase with distance (Levin, 2001). More generally, this pattern of results may be related to the faster decay of syntactic details of a sentence (such as morpho-syntactic agreement features) relative to the representation of its meaning (Sachs, 1967).

In the present study, across the two experiments, the conditions that led to the lowest relative acceptability rates were those in which an initial semantic (i.e. plural) agreement dependency was followed by a later syntactic (i.e. singular) agreement dependency, with the same agreement controller. This suggests that, once a collective agreement controller has been implicated in a semantic agreement process in incremental processing, it becomes difficult for later processes to access morpho-syntactic agreement for the same agreement controller. Conversely, the move from syntactic to semantic agreement, even within the same sentence, is more natural, and fits with the general pattern established for the semantic agreement preference to increase with distance across sentences (e.g. Schweppe, 2013).

Accounts that might explain why morpho-syntactic agreement becomes less accessible once semantic agreement has taken place will inevitably be speculative pending further research. One possibility is that the discourse representation of the referent of a collective noun like *government* may often be underspecified in terms of aspects of its interpretation—for example, whether it is interpreted as a single entity or as a collection of individuals. Further, a singular agreement target (e.g. the verb *was*) may be processed without adding interpretative detail to this underspecified representation, matching the collective noun phrase purely at the morpho-syntactic level. In contrast, according to this proposal, a plural agreement target, being the marked case, would be more likely to lead to an update in the discourse representation, signalling that the referent should be interpreted as a collection of individuals. Once this further specification has been made, there would be a preference for subsequent agreements involving the same agreement controller to match this semantic commitment. This would be consistent with the results reported here, since the relatively acceptable agreement mismatch cases were those in which a singular agreement target was followed by a plural agreement target. The initial singular agreement may have had a minimal effect on the semantic representation of the collective noun phrase, while the later plural agreement may have added extra detail. In contrast, the relatively unacceptable mismatch cases were those in which an initial plural agreement was followed by a subsequent singular agreement. According to this proposal, the initial plural agreement would lead to an update in the discourse representation, which would then clash with the later singular agreement. Assuming that such a clash affects acceptability ratings, this could explain the results.

Alternatively, it may be possible to overlay processing mechanisms on some of the theoretical linguistic accounts that we discussed above. For example, both den Dikken (2001) and Pollard and Sag (1994) require the verb and anaphor to match in number marking, for the types of sentences that we examine in this paper, at least at the competence level. However, if we assume

(a) that these competence grammars can be combined with incremental parsers, and (b) that the representation of the collective noun phrase may change during the processing of a single sentence, then mismatches in number could be captured. For example, in the case of den Dikken (2001), whether *the committee* participates in plural or singular agreement depends on whether or not an empty plural pronominal is adjoined to this phrase, while in the case of Pollard and Sag (1994), this depends on whether the referent is assigned a singular or a plural index. During the left-to-right processing of the sentence, this aspect of the representation could be updated. For example, in the case of den Dikken (2001), the plural empty pronominal could be inserted, presumably reflecting an update in the interpretation of the referent. This would mean that one could observe singular agreement at the first agreement target (e.g. the verb in Experiment 1) if this element is processed before the insertion of the pronominal, and subsequent plural agreement at the second agreement target (e.g. the anaphor in Experiment 1), if it is processed after the insertion, leading to a mismatch in number between verb and anaphor. In order to capture the results of this paper, one would need further assumptions in order to capture the asymmetry; for example, that it is more difficult to delete the empty pronominal than it is to insert it.

Above, we have discussed competence-based proposals, and how they may potentially be combined with performance-based mechanisms. However, instead of seeing competence and performance as two factors that have to be reconciled, another approach would be to consider models that do not cleanly make such a distinction. Self-Organized Sentence Processing (SOSP) models (G. Smith et al., 2018, 2021; Villata & Tabor, 2022) may be suitable for examining agreement mismatches. In these models, small fragments of structure known as *treelets* combine together to form larger structures, based on graded satisfaction of requirements between head and dependent treelets. G. Smith et al. (2018) show how such models can be applied to the investigation of notional number agreement in pseudo-partitive constructions like (10):

(10) A box of oranges *is/are*…

In their simulation of (10), G. Smith et al. (2018) allowed semantic features to influence whether the model would reach a stable state where either the singular *box* or the plural *oranges* was represented as the head of the phrase *a box of oranges*, thus predicting whether *is* or *are* would be chosen as the auxiliary.

The types of sentences examined in the current paper are similar in some ways to the pseudo-partitives examined by G. Smith et al. (2018), in that a noun phrase may participate in either singular or plural agreement, depending on semantic features. A successful model of double mismatches would be one where the model can reach a stable state with one dependency showing plural agreement while the other shows singular agreement, and then it would be possible to test the preferences for one type of mismatch over another.

5. Conclusion

We have described two experiments that examine the *double mismatch* effect in number agreement for morphologically singular collective noun phrases in English. The results show relatively high acceptability for sentences in which a collective noun controls singular agreement with an initial target, followed by plural agreement with a later target, while the opposite configuration results in lower acceptability. Given that both possible orderings of verb and anaphor were represented across the two experiments, this provides support for the *order-based account* as described above, suggesting that incremental processing will need to play an important role in any theoretical account. Finally, note that we are not claiming that relative distance between controllers and targets can provide a complete explanation for the semantic vs. syntactic agreement biases of the agreement hierarchy, or that the parts of speech of the agreement targets play no role. However, based on the results reported here, we argue that linear order plays a major role in explaining the double mismatch effect.

Supplementary materials

Stimuli, data and analysis scripts are available at: https://osf.io/zvsar/.

Acknowledgements

This research was supported by a PPLS Small Project Grant, awarded by the University of Edinburgh. This work was presented in 2022 at the HSP conference, the PROMS workshop and the Summer School of Linguistics 2022. I thank the participants of these events for interesting discussion. I would also like to thank three anonymous reviewers for very useful comments on the manuscript. All remaining errors are my own.

Competing interests

The author has no competing interests to declare.

References

Ackema, P., & Neeleman, A. (2018). *Features of person: From the inventory of persons to their morphological realization*. Linguistic Inquiry Monograph. DOI: https://doi.org/10.7551/mitpress/11145.001.0001

Akmajian, A. (1970). *Aspects of the grammar of focus in English* (Doctoral dissertation, MIT).

Bock, J. K., Butterfield, S., Cutler, A., Cutting, J. C., Eberhard, K. M., & Humphreys, K. R. (2006). Number agreement in British and American English: Disagreeing to agree collectively. *Language*, *82*(1), 64–113. DOI: https://doi.org/10.1353/lan.2006.0011

Bock, J. K., Eberhard, K. M., & Cutting, J. C. (2004). Producing number agreement: How pronouns equal verbs. *Journal of Memory and Language*, *51*, 251–278. DOI: https://doi.org/10.1016/j.jml.2004.04.005

Bock, J. K., Nicol, J., & Cutting, J. C. (1999). The ties that bind: Creating number agreement in speech. *Journal of Memory and Language*, *40*, 330–346. DOI: https://doi.org/10.1006/jmla.1998.2616

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. DOI: https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. DOI: https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. DOI: https://doi.org/10.18637/jss.v100.i05

Corbett, G. G. (1979). The agreement hierarchy. *Journal of Linguistics*, *15*, 203–224. DOI: https://doi.org/10.1017/S0022226700016352

Corbett, G. G. (1983). *Hierarchies, targets and controllers: Agreement patterns in Slavic*. London & Canberra: Croom Helm.

Corbett, G. G. (2000). *Number*. Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139164344

den Dikken, M. (2001). Pluringulars, pronouns and quirky agreement. *The Linguistic Review*, *18*(1), 19–41. DOI: https://doi.org/10.1515/tlir.18.1.19

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. DOI: https://doi.org/10.1016/j.jml.2013.04.003

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

Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition*, *101*(1), 173–216. DOI: https://doi.org/10.1016/j.cognition.2005.10.003

Furuya, K. (2017). (Under) specification of the person feature in relative clauses. *Acta Linguistica Academica*, *2*, 281–311. DOI: https://doi.org/10.1556/2062.2017.64.2.6

Godard, D. (1989). Empty categories as subjects of tensed Ss in English or French? *Linguistic Inquiry*, *20*(3), 497–506.

Haskell, T. R., & Macdonald, M. C. (2003). Conflicting cues and competition in subject-verb agreement. *Journal of Memory and Language*, *48*(4), 760–778. DOI: https://doi.org/10.1016/S0749-596X(03)00010-X

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. DOI: https://doi.org/10.1017/9781316423530

Kayne, R. S. (1994). *The antisymmetry of syntax*. Cambridge, MA: MIT Press.

Kreiner, H., Garrod, S., & Sturt, P. (2013). Number agreement in sentence comprehension: The relationship between grammatical and conceptual factors. *Language and Cognitive Processes*, *28*, 829–874. DOI: https://doi.org/10.1080/01690965.2012.667567

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. DOI: https://doi.org/10.7551/mitpress/6393.001.0001

Levin, M. (2001). *Agreement with collective nouns in English*. Lund Studies in English 103, M. Thormählen and B. Warren, eds. Lund: Lund University.

Molinaro, N., Kim, A., Vespignani, F., & Job, R. (2008). Anaphoric agreement violation: An ERP analysis of its interpretation. *Cognition*, *106*, 963–974. DOI: https://doi.org/10.1016/j.cognition.2007.03.006

Pearlmutter, N., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*(3), 427–456. DOI: https://doi.org/10.1006/jmla.1999.2653

Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. Stanford, CA.: CSLI and University of Chicago Press.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics*, *2*, 437–442. DOI: https://doi.org/10.3758/BF03208784

Sauerland, U., & Elbourne, P. (2002). Total reconstruction, PF movement, and derivational order. *Linguistic Inquiry*, *33*(2), 283–319. DOI: https://doi.org/10.1162/002438902317406722

Schweppe, J. (2013). Distance effects in number agreement. *Discourse Processes*, *50*(8), 531–556. DOI: https://doi.org/10.1080/0163853X.2013.841074

Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject-verb number agreement. *Cognitive Science*, *42*(4), 1043–1074. DOI: https://doi.org/10.1111/cogs.12591

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, *124*, 101356. DOI: https://doi.org/10.1016/j.cogpsych.2020.101356

Smith, P. (2015). *Feature mismatches: Consequences for syntax, morphology and semantics* (Doctoral dissertation, University of Connecticut, Storrs).

Smith, P. (2017). The syntax of semantic agreement in English. *Journal of Linguistics*, *53*, 823–863. DOI: https://doi.org/10.1017/S0022226716000360

Steele, S. (1978). Word order variation: A typological study. In J. Greenberg, C. Ferguson & E. Moravcsik (Eds.), *Universals of human language: IV: Syntax* (pp. 585–623). Stanford: CA: Stanford University Press.

Van Valin, R. D. (1986). An empty category as the subject of a tensed S in English. *Linguistic Inquiry*, *17*, 581–586.

Vergnaud, J.-R. (1974). *French relative clauses* (Doctoral dissertation, MIT, Cambridge: MA).

Villata, S., & Tabor, W. (2022). A self-organized sentence processing theory of gradience: The case of islands. *Cognition*, *222*, 104943. DOI: https://doi.org/10.1016/j.cognition.2021.104943

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

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