Agreement attraction in grammatical sentences and the role of the task

Anna Laurinavichyute*

Department of Linguistics, University of Potsdam, Haus 14, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany

Titus von der Malsburg

Institute of Linguistics, University of Stuttgart, Keplerstraße 17, 70174 Stuttgart, Germany

Author Note

*Corresponding author: Anna Laurinavichyute, anna.laurinavichyute@uni-potsdam.de

Abstract

This study evaluates two broad classes of language processing accounts that make predictions for sentences like "The admirer of the singer(s) apparently thinks...". Feature distortion accounts predict increased processing difficulty at the verb in sentences with a plural distractor noun (*singers*) while similarity-based interference accounts predict the opposite: increased difficulty in sentences with singular distractor noun (*singer*). Neither of these effects was reliably observed in earlier research, and the Bayesian meta-analysis of 31 published studies reported here is almost perfectly inconclusive. An explanation may be that both effects occur simultaneously and therefore mask each other. To test this idea, we conducted three single-trial self-paced reading experiments ($N_1 = 4,296$, $N_2 = 3,920$, $N_3 = 3,559$) which orthogonally manipulated agreement attraction and inhibitory interference. Surprisingly, all three experiments produced evidence for agreement attraction but none for inhibitory interference. Experiment 4 ($N_4 = 3,535$) tested the role of the expected task by preparing participants for a comprehension question (vs. acceptability judgment in Experiments 1–3). It showed neither agreement attraction nor inhibitory interference effects. Our findings demonstrate that agreement attraction effects can arise in grammatical sentences – contra earlier research – but also that these effects crucially depend on the task. This finding explains inconsistent results in prior research and it creates challenges, and opportunities, for future research.

Keywords: sentence comprehension, agreement attraction, illusion of ungrammaticality, similarity-based interference, adaptation, task effects

Agreement attraction in grammatical sentences and the role of the task

Introduction

Consider this pair of sentences:

(1) a. The computer installed in the old assembly station is unreliable.
b. The computer installed in the old assembly stations is unreliable.

Both sentences require the formation of a syntactic dependency between the verb "is" and its subject "computer". Even though both sentences are easy to understand, feature distortion accounts of *agreement attraction* in comprehension predict more processing effort at the verb when the number of the distractor noun "assembly station(s)" differs from the number of the subject noun. The particular mechanisms driving the increase in processing effort differ between accounts: For example, the singular subject noun phrase could be miscast as plural due to erroneous percolation of the plural feature of the distractor noun to the subject noun phrase (Bock & Eberhard, 1993; Eberhard et al., 2005). Alternatively, the configuration in (1-b) could create confusion about where the plural feature belongs (Dempsey et al., 2022; Konieczny et al., 2004). Both mechanisms would lead to slowdowns in the processing of the singular verb in (1-b) compared to (1-a), either because subject-verb agreement is perceived to be broken, or because it is difficult to establish the agreement.

Early feature distortion accounts were developed to explain agreement errors in production, which happen more often when a distractor noun mismatches the subject in number (particularly when the subject is singular). However, effects in comprehension largely mirror the agreement errors found in production: When a sentence with an agreement attraction error, such as (2), is encountered, the verb "are" is perceived to be more acceptable and is processed faster when there is a distractor noun that matches the verb in number (*illusion of grammaticality*, Hammerly et al., 2019; Patson & Husband, 2016; Wagers et al., 2009). Hence, it is

natural to consider production accounts also as an explanation for comprehension; and feature distortion accounts by Konieczny et al. and Dempsey et al. explicitly cover agreement attraction effects in comprehension.

(2) a. The computer installed in the old assembly station *are ...

b. The computer installed in the old assembly stations *are ...

Feature distortion accounts also make predictions about the processing of grammatical sentences. Due to the distractor noun that mismatches the verb in number, (1-b) should be perceived to be less acceptable than (1-a) even though both are equally grammatical – the so-called *illusion of ungrammaticality* (Wagers, 2008). Likewise we would expect slower reading times in (1-b) compared to (1-a) (Nicol et al., 1997; Pearlmutter et al., 1999). However, none of these effects have been consistently attested, which casts doubt on the ability of the feature distortion accounts to fully explain comprehension.

A recent meta-analysis estimated the speedup observed in ungrammatical sentences with attraction, (2-b) vs. (2-a), to be -22 ms (95% credible interval: $[-36, -9]$ ms, Jäger et al., 2017). However, the evidence for the predicted slowdown in grammatical sentences, (1-b) vs. (1-a), is much less clear. While several studies reported such slowdowns, most of them had design shortcomings. One widely acknowledged problem is spillover from the sentence region containing the distractor noun which is read more slowly when the noun is plural. Most other studies found no difference at all. Jäger et al. estimated the effect to be 7 ms but with a 95% credible interval (95%-CrI) of $[-4, 16]$ ms which allows for a positive, a negative, as well as no effect. Our own meta-analysis which included more recent results estimated the effect to be 1 ms with a 95%-CrI of $[-7, 9]$ ms (see section "Meta-analysis" below). The literature therefore offers no basis for the idea that the number of a distractor noun has an impact on the processing of the subject-verb dependency in grammatical sentences like (1).

Interestingly, sentences such as (1) present a problem not only for feature

distortion accounts. Similarity-based interference accounts (Lewis & Vasishth, 2005; McElree, 2000), too, make predictions for these sentences. They posit that the formation of the subject-verb dependency is harder and takes more time when the subject shares morphosyntactic features with the distractor noun. As a result, (1-a) should be harder to process than (1-b). This prediction is, of course, the exact mirror image of the effect predicted by the feature distortion accounts.

The explanation of the slowdown in (1-a) vs. (1-b) is as follows: Similarity-based interference accounts assume that ongoing sentence comprehension relies on a series of fast retrievals of previously processed constituents from content-addressable memory (Lewis & Vasishth, 2005; Lewis et al., 2006; McElree, 2000; Van Dyke & McElree, 2006). The speed and accuracy of these retrievals depend on how unique the features of the to-be-retrieved element are. If the to-be-retrieved element ("computer") shares features, in this case number, with other elements in memory ("assembly station"), retrieval will on average take longer. The resulting slowdown in (1-a) compared to (1-b) is called *inhibitory interference*.

Empirical support for inhibitory interference in number is rather weak: The predicted effect was found in only two studies (Franck et al., 2015; Nicenboim et al., 2018). Most other studies were inconclusive as reflected in the meta-analysis by Jäger et al., which shows that conditions like (1-a) are read faster, not slower than (1-b) by 7 ms (95%-CrI: $[-4, 16]$ ms).

To summarize, feature distortion and similarity-based interference accounts predict opposite effects in the processing of grammatical sentences such as (1). Neither of these effects has been consistently observed, and the outcomes of meta-analyses were inconclusive, i.e. compatible with either account or the lack of any effect. These inconsistent results are equally problematic for both groups of accounts. They restrict the scope of feature distortion accounts to language production and the comprehension of ungrammatical sentences. At the same time these results cast doubt on the existence of morphosyntactic inhibitory interference in grammatical sentences. Morphosyntactic interference is a key component in

prominent models of sentence comprehension, such as the Lewis and Vasishth model (Lewis & Vasishth, 2005) or the direct access model (McElree, 2000), and the absence of interference in grammatical sentences therefore challenges core assumptions of these models.

In the following sections, we will review the evidence collected to date and summarize it in a new meta-analysis. We will then propose that agreement attraction and inhibitory interference effects may be present simultaneously and that no consistent picture has emerged because these effects partially mask each other in the configurations commonly used for testing. If true, the inconsistency in the prior studies would be explained and the case for both feature distortion and similarity-based interference accounts would be considerably strengthened.

To test the masking hypothesis, we conducted three preregistered self-paced reading experiments that aimed to differentiate the effects of feature distortion and inhibitory interference. To preview the results: Contrary to our hypothesis, only effects consistent with feature distortion (i.e., agreement attraction effects) were observed. These results stand in striking contrast to previous research, and we explore the causes of this discrepancy in Experiment 4, which suggests that the task (comprehension questions vs. acceptability judgments using in Experiments 1–3) may play a more important role than previously thought.

Meta-analysis

The predictions of similarity-based interference and feature distortion accounts for sentences such as (1-a) and (1-b) contradict each other: similarity-based interference predicts a relative slowdown in (1-a) compared to (1-b) whereas feature distortion predicts the opposite.

Only two studies reported an effect compatible with inhibitory interference (Franck et al., 2015; Nicenboim et al., 2018). On the other hand, a number of studies reported slowdowns compatible with feature distortion (Pearlmutter et al., 1999; Nicol et al., 1997, Expts. 1, 2, 4, 5; Wagers et al., 2009, Expt. 4; Lago et al.,

2015, Expt. 3A; Patson and Husband, 2016, Expt. 1; Smith et al., 2021, Expt. 1). Unfortunately, many of these studies had design confounds: the slowdown detected at the verb might be a spillover effect from the previous region, the plural noun, and therefore reflect unrelated processes. Plural nouns might take more time to process than singular nouns for several reasons: They are longer, less frequent, morphologically more complex, and might be more difficult to integrate into the discourse. Although the slowdowns detected in those studies cannot unequivocally support the feature distortion accounts, similar slowdowns have also been reported in a study that used the maze-task paradigm, which is claimed to be largely free from spillover effects (Expts. 1, 2, and 4 by Nicol et al., 1997; for an evaluation of the maze task, see Boyce et al., 2020).

Reading studies that showed no slowdown on the region preceding the verb, and which were therefore likely free from the spillover confound, found only weak evidence for attraction (Smith et al., 2021, Expt. 1; Lago et al., 2015, Expt. 3A) or no effect at all (Nicol et al., 1997, Exp. 3; Slioussar, 2018, Expt. 3; Wagers et al., 2009, Expts. 1-3, 5; Thornton and MacDonald, 2003, Expt. 3; Patson and Husband, 2016, Expt. 2; Villata and Franck, 2020, Expt. 3; Lago et al., 2015, Expts. 1, 2, 3B; Avetisyan et al., 2020, Expts. 2, 3; Brehm et al., 2019; Dempsey et al., 2022; Jäger et al., 2020; Lago et al., 2021; Paape et al., 2021; Parker and An, 2018; Tucker et al., 2015; Smith et al., 2021, Expt. 2). In sum, feature distortion predicts a slowdown in (1-b) compared to (1-a), but this effect is rarely observed in experiments without design limitations.

The absence of the agreement attraction effect is widely viewed as a failure of feature distortion accounts to explain comprehension in grammatical sentences. However, this view has recently been challenged by studies using other dependent measures. For instance, Hammerly et al. (2019) argue that the illusion of ungrammaticality is present in grammaticality judgments, but masked by a bias to the "grammatical" response. When response bias was neutralized, the illusion of ungrammaticality surfaced. Similarly, a visual world study by Brehm et al. (2021)

demonstrated that after having heard "The key to the cabinets..." but before hearing the verb, participants look at the image of several keys more often than after having heard "The key to the cabinet...". Both findings support feature distortion accounts, but do not shed light on the lack of the predicted slowdowns in reading times.

In order to summarize and evaluate the accumulated evidence, we conducted a meta-analysis that included 16 estimates from the previous meta-analysis by Jäger et al. (2017) and 15 new estimates from studies published more recently: Avetisyan et al., 2020; Brehm et al., 2019; Jäger et al., 2020; Lago et al., 2021; Nicenboim et al., 2018; Paape et al., 2021; Parker and An, 2018; Slioussar, 2018; Smith et al., 2021; Villata and Franck, 2020.[1] Some theoretically relevant studies could not be included in this meta-analysis: Two experiments reported by Patson and Husband (2016) were not included because the effect on the millisecond scale could not be inferred from the reported residual log reading times. A recent study by Dempsey et al. (2022) was also not included because it did not target the subject-verb dependency, and therefore the predictions of the similarity-based interference accounts could not be evaluated.

The meta-analysis model consisted of an intercept (an overall meta-analytic estimate) and nested random effects: by-publication and by-experiment random intercepts, with experiments nested within publications. This structure can capture potentially higher similarities between the experiments reported within a single publication, where participants are drawn from the same pool, data trimming and analyses are conducted by the same investigator, etc. The details of model fitting, including the specification of priors, can be found in Appendix B.

The results of the meta-analysis are presented in Figure 1. Compared to the meta-analysis by Jäger et al., who reported an effect of 7 ms, 95% CrI $= [-4, 16]$ ms, the new estimate is closer to zero and spans negative and positive values almost equally ($\hat{\beta} = 1$ ms, 95%-CrI: $[-7.2, 9]$ ms). Thus, the new meta-analysis does not

[1] Although the relevant comparison is not directly reported in Brehm et al. (2019) and Villata and Franck (2020), the authors shared the estimates in personal communication.

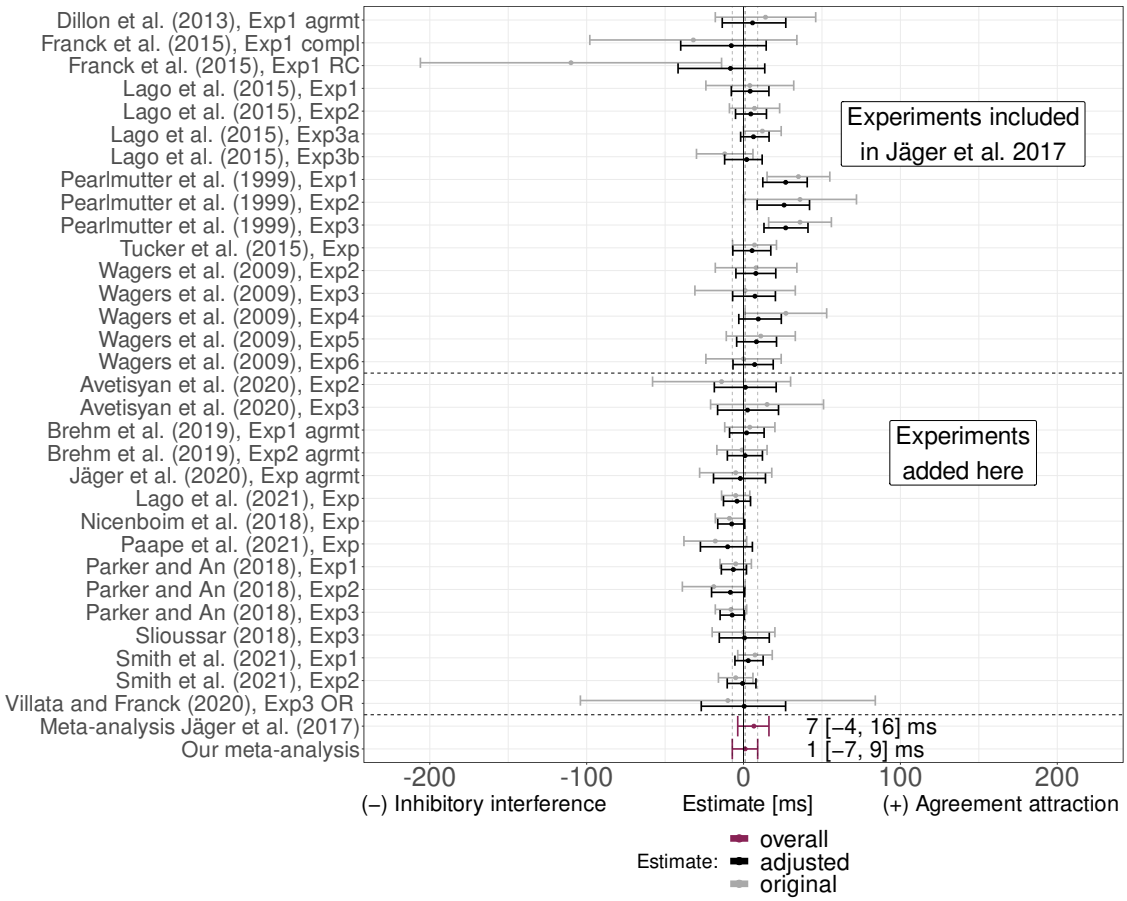

Figure 1. Quantitative summary of previous findings and the results of the meta-analysis. The estimates reported in the individual studies are in light gray, the adjusted estimates computed in the meta-analysis are in black. The meta-analytic estimates of the current and previous analyses are in purple.

lend support to either of the accounts.

The masking hypothesis

Given that more than 30 studies conducted over more than 20 years did not render a conclusive picture, the issue may lie not in a lack of data, but on the theoretical side. Agreement attraction and inhibitory interference effects are the opposite of each other but, crucially, they need not be mutually exclusive. We therefore need to consider the possibility that feature distortion and inhibitory interference may operate simultaneously, and that the resulting effects may mask each other thus giving the false impression that neither agreement attraction effects nor interference effects arise in grammatical sentences.

However, we cannot simply assume that both mechanisms operate

independently. In fact, they cannot: Feature distortion accounts assume that morphosyntactic features are occasionally represented incorrectly due to either feature percolation or uncertainty as to where they belong. Incorrect representation of morphosyntactic features should immediately affect parsing as conceptualized in similarity-based interference. We therefore need to carefully unpack how both mechanisms could operate together in order to determine what reading time patterns should be expected in a combined account.

First, we note that feature distortion must occur before the retrieval of the subject at the verb. In configurations such as (1-b), feature distortion will affect only some trial in which the parse will then represent the ungrammatical (and non-veridical) sentence "The *computers* installed in the old *assembly stations* is ...". According to similarity-based interference, ungrammatical structures are processed more slowly than their grammatical counterparts because there is no perfect match for the verb's retrieval cues (Lewis & Vasishth, 2005). The processing times for configurations such as (1-b) will therefore consist of a mixture of short retrieval times for correct parses and long retrieval times for incorrect parses resulting from feature distortion (see Figure 2b), which means that the average processing time in (1-b) will become slower and can approach that of (1-a) where a slowdown due to inhibitory interference is expected.

The extent of the average slowdown in (1-b) depends on the rate at which feature distortion happens. In the example in Figure 2, we assume a rate of 20%, roughly the average of the rates reported in the literature: 17% in Schlueter et al. (2019), 18% in Staub (2009), 13% and 19% in Thornton and MacDonald (2003), and 21%, 6%, and 29% in Laurinavichyute and von der Malsburg (2022) (a recent meta-analysis by Yadav et al. reports an estimate of 12%–22%). Even with lower feature distortion rates, the difference in reading times between (1-b) and (1-a) will be diminished and will become difficult to detect even if it is present.

All in all, these considerations show the following: Even though feature distortion and inhibitory interference interact in a slightly non-additive way, the

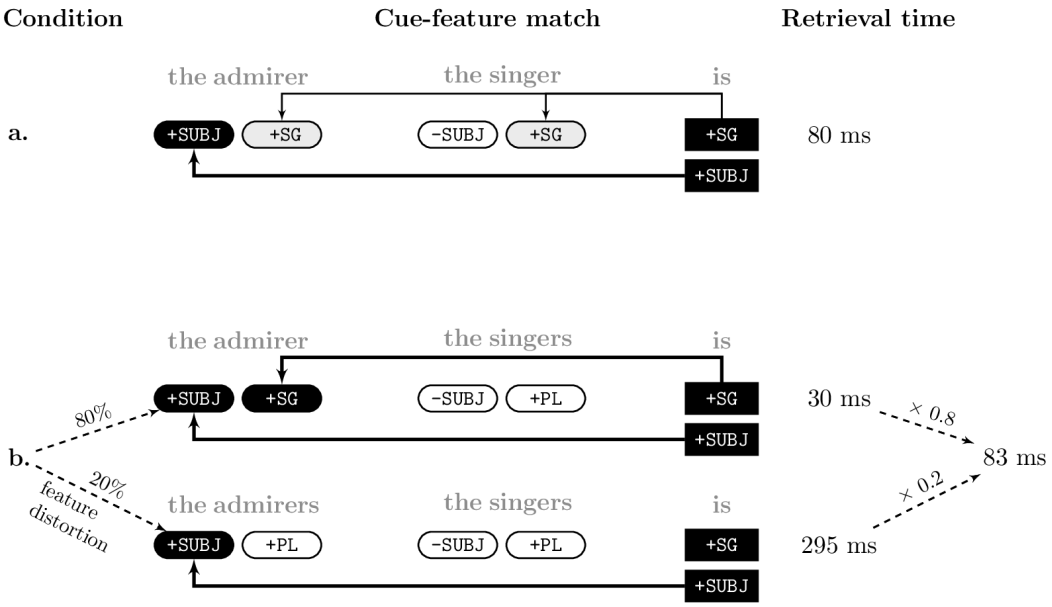

Figure 2. Schematic illustration of how feature distortion and inhibitory interference could co-exist and lead to similar processing times in grammatical sentences with singular and plural distractor nouns. The estimates of retrieval times were generated using https://engelmann.shinyapps.io/inter-act with settings: Interference type: `retroactive interference` and Model: `Classic LV05 model`.

overall idea, namely that their effects could largely mask each other, is plausible. Recent computational modeling work by Yadav et al. (2023) (partly inspired by an earlier version of the present study, Laurinavichyute, 2021) also suggests that simultaneous deployment of both feature distortion and similarity-based interference mechanisms provides a better fit to several published data sets than any single mechanism operating alone.

The fact that the modeling results are consistent with the masking hypothesis is encouraging but does not unequivocally support the masking hypothesis (see Yadav et al. for possible alternative explanations). Only a direct test can show that both inhibitory interference and feature distortion are at play.

One way to test whether both interference and feature distortion mechanisms simultaneously affect processing times is to manipulate inhibitory interference independently from feature distortion. Consider Example (3-a): Here, inhibitory interference consists of two components. The distractor noun “singer” not only

shares the number marking of the verb "thinks", but is also a plausible theme of the verb, which induces semantic interference (Van Dyke, 2007; Van Dyke & McElree, 2006, 2011). To manipulate overall interference in (1-a), it would therefore suffice to make the distractor noun semantically less compatible with the verb.

The experimental conditions implementing this idea are illustrated in Example (3): (3-a) and (3-b) mirror the traditionally tested conditions where the distractor noun matches the verb in number and thematic requirements. In contrast, in conditions (3-c) and (3-d) the distractor noun is inanimate and does not meet the thematic requirements of the verb.[2]

(3) a. The admirer of the singer apparently thinks ...
b. The admirer of the singers apparently thinks ...
c. The admirer of the play apparently thinks ...
d. The admirer of the plays apparently thinks ...
... the show was a big success.

To obtain quantitative model predictions for these configurations, we modified the feature percolation-plus-retrieval model by Yadav et al. (2023) to work with three features (syntactic, semantic, and number). All parameter values, including the possible range of feature distortion rates, were kept the same as in Yadav et al. (2023), except for the mismatch penalty. Mismatch penalty determines the decrease in activation (and therefore, the slowdown in retrieval times) for an element in memory that matches some but not all retrieval cues. It comes into play in those trials where feature distortion occurs and where the subject noun is no longer the perfect match for the verb. The mismatch penalty used by Yadav et al. (2023), 0.15, was estimated by Engelmann et al. (2019). However, this estimate is based mostly on studies with many ungrammatical stimuli, and participants' sensitivity to

[2] Notice that the design is similar (though not identical) to the design of grammatical conditions reported in Thornton and MacDonald (2003). However, the results reported by Thornton and MacDonald cannot be evaluated since the interaction critical for our argument was not reported and cannot be inferred from their plots.

feature mismatch might have been decreased. The default mismatch penalty value in ACT-R is 1 (Anderson, 1996), and we present the simulation results for both values (see Figure 3). Model predictions are based on mean reading times for each of four conditions for 1000 simulated data sets.[3]

If both feature distortion and similarity-based interference simultaneously affect reading times, two possible outcomes are expected, depending on the value of the mismatch penalty parameter. If mismatch penalty is low, only inhibitory semantic interference of 11 ms, 95% CrI = [5, 18] ms is expected, other effect estimates are centered around 0; if mismatch penalty is high, three effects are expected: inhibitory semantic interference of 16 ms, 95% CrI = [6, 28] ms, agreement attraction effect of 33 ms, 95% CrI = [−6, 93] ms, and an interaction between these fixed effects, 8 ms, 95% CrI = [−2, 23] ms. Therefore, a pattern of condition means that fits either simulation would directly support the masking hypothesis; in addition, it would speak to the likely value of the mismatch penalty parameter in the processing of grammatical sentences.

Data availability statement

All materials (experimental items and comprehension questions, training sentences, and instructions), data, and analysis code used for this paper, including the meta-analysis and simulations, are openly available online at https://osf.io/vyqe3/?view_only=0ec02ba90b264531a7ccda407132b97c. The anonymized link will be replaced with a permanent link or DOI after peer review.

The hypotheses, number of participants, and analyses planned for Experiment 1 were pre-registered on OSF (doi:10.17605/OSF.IO/PD8KY). The hypotheses, number of participants, and analyses planned for Experiments 2 and 3 were pre-registered separately on OSF (doi:10.17605/OSF.IO/VM5BW, on the left side menu, go to Wiki → Component Wiki pages).

[3] The pre-registered hypothesis is different from the simulations presented here as the pre-registered hypothesis was based on faulty assumptions about how the model should behave instead of direct simulations. The authors are grateful to Christopher Hammerly for pointing this out to them.

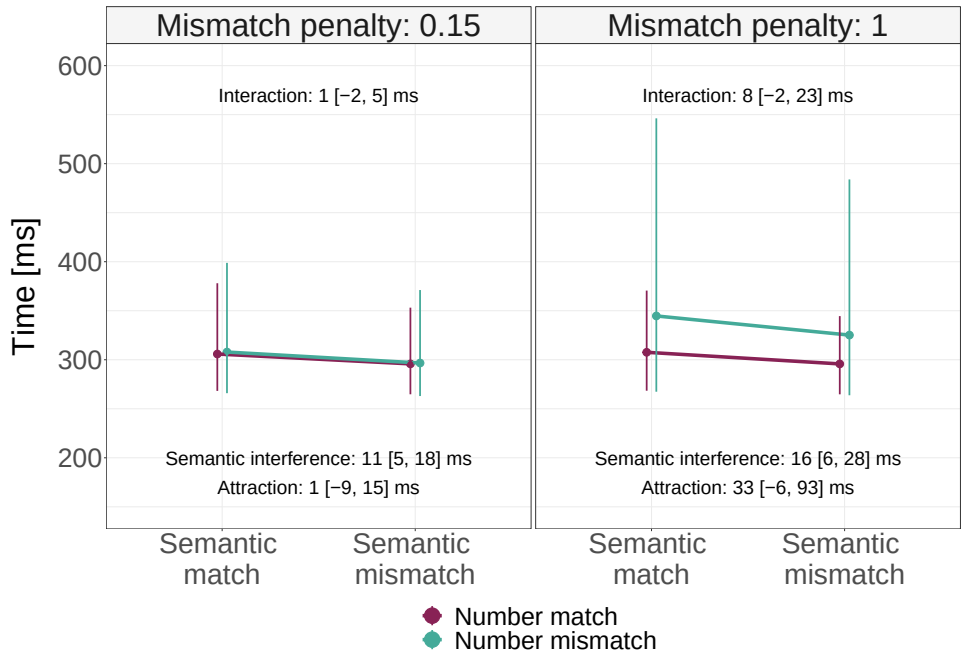

Figure 3. The patterns of reading times predicted under both inhibitory interference and feature distortion, depending on the mismatch penalty parameter. The error bars represent the whole range of simulated condition means.

Experiment 1

The aim of Experiment 1 was to test whether the inhibitory interference effect predicted by similarity-based interference and the agreement attraction effect predicted by the feature distortion accounts are simultaneously present in the reading times for grammatical sentences. To do that, we manipulated similarity-based interference independently from feature distortion as explained above.

The hypotheses, number of participants, and analyses planned for Experiment 1 were pre-registered on OSF, doi:10.17605/OSF.IO/PD8KY. Deviations from the pre-registered analysis are explained below.

Methods

Participants. Participants were recruited on the crowdsourcing platform Prolific and were compensated based on the recommended hourly rate of 6£ per hour. A compensation of 10 pence was offered for the task of reading and rating

four sentences, which took approximately one minute. Inclusion criteria for participants were: (i) being a native speaker of English and (ii) being a resident of the US, UK, Ireland, New Zealand, or Australia.

Based on the power calculations described in the preregistration, we estimated that 4,160 participants (65 independent observations per item and condition) would ensure a statistical power between 61% and 88%, depending on the effect size, ranging from 0.017 to 0.025 log milliseconds which corresponds approximately to an 2% increase/decrease of reading times (e.g., 8 ms when the RT in the control condition is 400 ms). However, in order to ensure sufficient data after application of exclusion criteria (below), we collected data from overall 5,577 participants.

We excluded data from participants who: (a) indicated in a post-experiment questionnaire that English was not their native language or that they do not currently live in an English-speaking country; (b) gave exactly the same rating to the three practice sentences (two well-formed sentences and one sentence with an apparent agreement error); (c) had reading times for any word in the experimental sentence that fell below 180 ms or above 3,000 ms. After applying these criteria, 4,296 participants were left for the analysis.

Materials. We created 16 items similar to (3) in a 2×2 design manipulating semantic and number match/mismatch between the distractor noun and the verb (the subject noun always fully matched the verb both in number and semantics). In the semantic match conditions (a) and (b), both nouns could potentially perform the action denoted by the verb. The distractor noun never referred to a multitude (such as “team”, “collective”, etc.). Within the sentence, the noun phrase was followed by an adverb and then a verb with correct number marking, the same across all conditions. The verb was followed by a region that was the same across conditions and did not indicate the number of the head noun (no personal pronouns, etc.).

Each item was followed by a comprehension question with five response options targeting the subject-verb dependency, as in Example (4). The question

rephrased the sentence and contained a verb marked for past simple tense, so that the verb provided no information about the number of the head noun. The response options were: the head noun in singular and plural forms, the distractor noun in singular and plural forms, and *I'm not sure*, presented in random order.

(4) Who considered the show a success? — Admirer/Admirers/Singer/Singers /I'm not sure. Or, in the semantic match conditions: Admirer/Admirers/Play/Plays/I'm not sure.

Item norming

To ensure that semantic match/mismatch was perceived as such by native English speakers, a plausibility norming study was conducted. Based on each item, three sentences were created whose subjects were the head noun, the animate distractor noun, and the inanimate distractor noun of the original item, respectively, see Example (5):

(5) a. The admirer of the play supposedly thinks [the show was a big success].
b. The singer supposedly thinks [the show was a big success].
c. The play supposedly thinks [the show was a big success].

We conducted two online questionnaires, both prompted participants to rate sentences on a ordinal scale from 1 (bad, unnatural) to 7 (good, perfectly natural). In the first questionnaire, full sentences were presented; in the second questionnaire, sentences were truncated after the main verb (the truncated part is denoted by square brackets in Example (5)). Truncated sentences were tested to ensure that the mismatch between the distractor noun and the verb was apparent right at the verb and not only later in the sentence based on non-verb cues.

A total of 277 individuals took part in the norming study: 179 participants saw full sentences, and 98 other participants saw truncated sentences. Items were presented using a Latin square design, i.e., each participant saw every item in just

one out of three conditions. The results of both norming studies confirmed that sentences with animate subjects, such as (5-a) and (5-b), received similarly high ratings (full sentences: 5.56 (SE=0.16) for condition (a), 5.53 (SE=0.16) for condition (b); truncated sentences: 5.14 (SE=0.18) for condition (a), 5.40 (SE=0.20) for condition (b)). Sentences with inanimate subjects, such as (5-c), received lower ratings, as desired (full sentences: 3.52 (SE=0.25); truncated sentences: 3.27 (SE=0.20)). Further details on the norming study, including the statistical analysis, can be found in Appendix A.

Procedure

The experiment was programmed using the Ibex[4] software and run on the IbexFarm cloud service. We used single-trial procedure: Each participant first saw the instructions[5], then three practice sentences to get used to the non-cumulative centered self-paced reading procedure, and then a single experimental sentence in one of the four conditions. This way, participants could not get used to the manipulation and could not develop experiment-specific processing strategies (for a more detailed discussion of single-trial experimental designs, see Laurinavichyute and von der Malsburg, 2022). For each sentence, including the practice sentences, acceptability ratings on the scale from 1 (bad) to 7 (good) were collected to ensure that participants paid attention to the task, and to get an offline measure for the illusion of ungrammaticality and interference effects. For the experimental sentence only, the acceptability rating task was followed by a comprehension question targeting the subject-verb dependency.

---

[4] http://spellout.net/ibexfarm

[5] "We will ask you to read four sentences word by word and rate them. Please read carefully because you will not be able to revisit earlier words. One of the sentences will be followed by a comprehension question. First, you will see a dash. To read the first word, press spacebar, then press it again to read the second word, and so on. Each sentence will be followed by a 1 to 7 rating scale: if you think that the sentence was rather well-formed, give it a higher rating, if it is not so well-formed or has apparent problems, give it a lower rating."

Analyses

All analyses were conducted with the R system for statistical computing (R Development Core Team, 2009). Data were analyzed using generalized linear mixed models fit in the Bayesian framework (Vasishth et al., 2018) using the R package 'brms' (Bürkner et al., 2017), which is a front-end for 'Stan' (Carpenter et al., 2017), a statistical system for full Bayesian inference. Plots were produced with the 'ggplot2' and 'tidybayes' packages (Kay, 2019; Wickham, 2016). The results are reported in terms of the posterior mean, the 95% percentile intervals (95%-CrI), and the posterior probability of the parameter in question being being greater than zero ($P(\beta > 0)$). Inferences were based on the latter quantity which corresponds straightforwardly to a Bayes factor testing effect direction (they are identical if prior odds of 1:1 are assumed, see Rouder et al., 2018; Tendeiro & Kiers, 2022). The cutoff value was specified as 0.975.[6]

Every model included the main effects of number and semantic match/mismatch and their interaction: number mismatch was coded as 1, number match as -1, such that a positive main effect corresponded to the illusion of ungrammaticality that we expected to find; semantic match was coded as 1, mismatch as -1, such that a positive main effect corresponded to inhibitory semantic interference. The models also had by-item random intercepts and slopes for the main effects and their interaction. By-subject random effects were not needed since each subject contributed only one measurement.

For the reading time analysis, a mixed-effects regression model assuming log-normal distribution of the residuals was fit. Mixture modeling was pre-registered as the second analysis step but not performed since the prerequisite for fitting the model (finding both attraction and interference effects) was not met. For the analysis of acceptability ratings, ordered logistic mixed-effects regression

[6] Although the pre-registration specified the cutoff value of 0.95, we later decided to follow a stricter criterion: The cutoff value of 0.95 corresponds to the alpha level of a one-sided t-test, which is known to be prone to type I errors, while the cutoff value of 0.975 that we choose here corresponds to the critical level of the two-sided t-test.

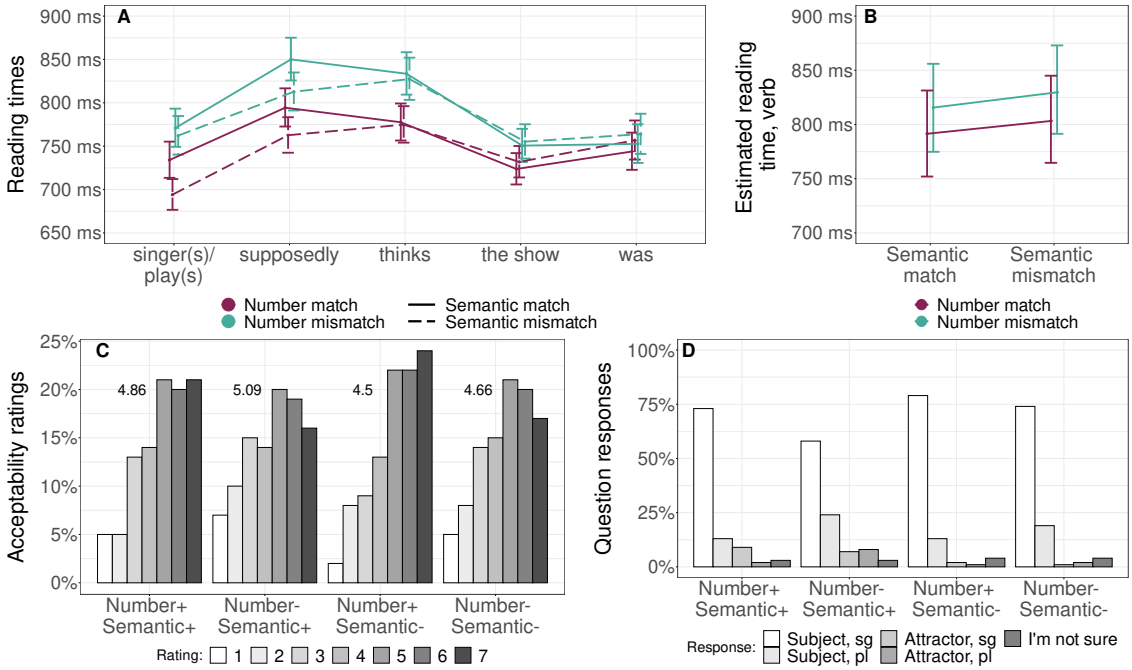

Figure 4. Results of Experiment 1. Panel A: reading times (geometric means) across sentence regions and 95% confidence intervals. Panel B: Estimated reading times at the verb with 95% credible intervals (spillover from the previous region is accounted for by statistical control). Panel C: proportions of acceptability ratings across conditions. Numbers represent mean rating in each condition. Panel D: proportions of question responses across conditions. In panels C and D, the x-axis encodes experimental conditions: *Number+* stands for number match, *Number−* for number mismatch; similarly, *Semantic+* stands for semantic match, and *Semantic−* for semantic mismatch.

models were used (Liddell & Kruschke, 2018; Veríssimo, 2021). Comprehension questions had five response options. In the descriptive statistics, proportions of responses of every category are presented for each condition, but for statistical analysis, we coded responses just as correct or incorrect. These binary responses were analyzed using logistic regression. The details of Bayesian model fitting, including the specification of regularizing priors, can be found in Appendix B.

Results

Summaries of reading times, acceptability ratings, and question response accuracies are presented in Figure 4.

Reading times. As can be seen from Figure 4A, an unexpectedly long-lasting plural complexity effect (a slowdown in the number mismatch conditions) spanned

three words following the plural distractor noun: the adverb, the critical verb, and the region following the verb. The slowdown on the pre-critical region renders the pre-registered analysis of reading times at the verb uninformative. We controlled the spillover effects by including the centered reading times from the previous word as a predictor for the reading times at the target word (Vasishth, 2006). This analysis shows whether the current word introduces any additional difficulties over and above those inherited from the previous word.

After applying the correction, a slowdown in the number mismatch conditions was found on the distractor noun itself, but crucially, not on the following adverb, suggesting that the correction worked as intended. At the verb, a main effect of number mismatch was again found: the verb was read more slowly in the condition with plural distractor nouns, consistent with the predictions of feature distortion accounts ($\hat{\beta} = 25\,\text{ms}$, 95%-CrI: $[0.67, 51]\,\text{ms}$, $P(\beta > 0) = 0.98$). There was no evidence for an effect of semantic match ($\hat{\beta} = -13\,\text{ms}$, 95%-CrI: $[-36, 9.8]\,\text{ms}$, $P(\beta < 0) = 0.89$). Likewise there was no evidence for an interaction ($\hat{\beta} = -1.2\,\text{ms}$, 95%-CrI: $[-24, 21]\,\text{ms}$, $P(\beta < 0) = 0.55$).

Acceptability ratings. Acceptability ratings were lower in the number mismatch condition ($\hat{\beta} = -9.4\%$, 95%-CrI: $[-14, -4.3]\%$, $P(\beta < 0) > 0.99$). There was a tendency toward lower ratings for the semantic match conditions ($\hat{\beta} = -5\%$, 95%-CrI: $[-11, 0.6]\%$, $P(\beta < 0) = 0.96$). There was no evidence for an interaction ($\hat{\beta} = 1.1\%$, 95%-CrI: $[-2.3, 4.5]\%$, $P(\beta > 0) = 0.75$).

Question response accuracies. Accuracy was lower in the number mismatch condition ($\hat{\beta} = -9.5\%$, 95%-CrI: $[-13, -6]\%$, $P(\beta < 0) > 0.99$) and in the semantic match condition ($\hat{\beta} = -11\%$, 95%-CrI: $[-16, -5]\%$, $P(\beta < 0) > 0.99$). There was also an interaction between the effects ($\hat{\beta} = -3.4\%$, 95%-CrI: $[-6.5, -0.18]\%$, $P(\beta < 0) = 0.98$). Nested comparisons show that the decrease in accuracy due to number mismatch was greater for semantic match ($\hat{\beta} = -14\%$, 95%-CrI: $[-19, -9]\%$) than for semantic mismatch ($\hat{\beta} = -5\%$, 95%-CrI: $[-10, -1]\%$).

Discussion

The outcomes of the experiment support feature distortion accounts. Feature distortion should lead to both the slowdown at the verb and the illusion of ungrammaticality (lower acceptability ratings) in number mismatch conditions. Surprisingly, there was no reading time or judgement evidence for inhibitory semantic interference or an interaction between number and semantic match. The absence of semantic interference goes against the masking hypothesis. We will address this in the general discussion below.

Some evidence supporting semantic interference comes from the question response accuracies. Although both number mismatch and semantic match decreased accuracy, these main effects are trivial: In both cases, the number of plausible response options is greater than in the conditions they are contrasted with. For example, in the number match conditions, two responses marked for plural are not viable as there are no plural words in the sentence, which is sufficient to account for higher accuracy. However, the interaction between the main effects may be more informative. The decrease in accuracy due to number mismatch was more pronounced within semantic match than within semantic mismatch conditions. This pattern is compatible with semantic but, importantly, not with number interference. Generally, the present results are in line with several studies reporting interference effects in grammatical sentences only in question responses, but not in reading times (Jäger et al., 2015; Laurinavichyute et al., 2017; Mertzen et al., 2020).

An unexpectedly long-lasting plural complexity effect spanning three regions rendered the planned analyses of reading times uninformative. The persistence of this effect is surprising since we used a typical design that takes the standard one-word spillover effects into account. Similar designs were used in many previous studies, and long-lasting plural complexity effects were reported only in Experiment 4 by Wagers et al. (2009). The single trial procedure might have enhanced the plural complexity effect: All effects, including those driven by confounding factors, are likely to be greater in the absence of adaptation to task and stimuli. This is a

potential strength of the single-trial procedure (high statistical power), but it is important to keep this feature in mind when designing materials. See Laurinavichyute and von der Malsburg (2022) for further discussion of single-trial designs.

To summarize, a slowdown on the verb in the number mismatch condition is compatible only with the feature distortion accounts, but this conclusion could be compromised by the statistical correction for spillover effects. Evidence from an experiment where the plural complexity effect is not present at the pre-critical region would be more convincing. We therefore conducted Experiments 2 and 3. In Experiment 2, we retain the materials from Experiment 1 but introduce a long parenthetical phrase between the distractor noun and the verb. In Experiment 3, we employ sentences with object relative clauses, where the distractor noun is located at a larger distance from the verb and its subject both linearly and structurally.

Experiment 2

Experiment 2 introduces a long parenthetical phrase between the distractor noun and the verb in order to mitigate the long-lasting plural complexity effect found in Experiment 1. Procedure and analysis are the same as in Experiment 1, except for the differences in the number of participants and experimental materials described below.

The hypotheses, number of participants, and analyses planned for Experiment 2 were pre-registered on OSF (doi:10.17605/OSF.IO/VM5BW; on the left side menu, go to Wiki → Component Wiki pages). Due to funding constraints, we pre-registered N=1,956 for each of Experiments 2 and 3. When additional funds became available later, we increased the number of participants to reach the number of observations estimated in the power analysis.

Participants

Participant recruitment and exclusion procedure were the same as in Experiment 1. We recruited only individuals who did not take part in Experiment 1

and received 4,831 responses. After applying exclusion criteria, data from 3,920 participants were left.

Materials

Materials from Experiment 1 were modified such that the distractor noun and the verb were separated by a parenthetical phrase three to five words long, see Example (6). The parenthetical contained either personal pronouns (*I*, *you*) or proper nouns (*Daily Mail*), but very few common nouns in order to keep additional interference as low as possible (Gordon et al., 2001). The parenthetical phrase was followed by the adverbial used in Experiment 1. In total, the buffer region between the distractor noun and the verb comprised five to seven words (5.5 on average).

(6) a. The admirer of the singer, according to the Daily Mail, apparently thinks...

b. The admirer of the singers, according to the Daily Mail, apparently thinks...

c. The admirer of the play, according to the Daily Mail, apparently thinks...

d. The admirer of the plays, according to the Daily Mail, apparently thinks...

... the show was a big success.

Results

A summary of reading times, acceptability ratings, and question response accuracies is presented in Figure 5.

Reading times. Introducing parenthetical phrases successfully eliminated the spill-over present in Experiment 1: No effects were observed on the four words preceding the critical verb. We therefore proceeded with the pre-registered analyses. No effects were found on the verb (number mismatch: $\hat{\beta} = 1.5\,\text{ms}$, 95%-CrI: $[-16, 21]\,\text{ms}$, $P(\beta > 0) = 0.56$; semantic match: $\hat{\beta} = 1.9\,\text{ms}$, 95%-CrI: $[-17, 21]\,\text{ms}$,

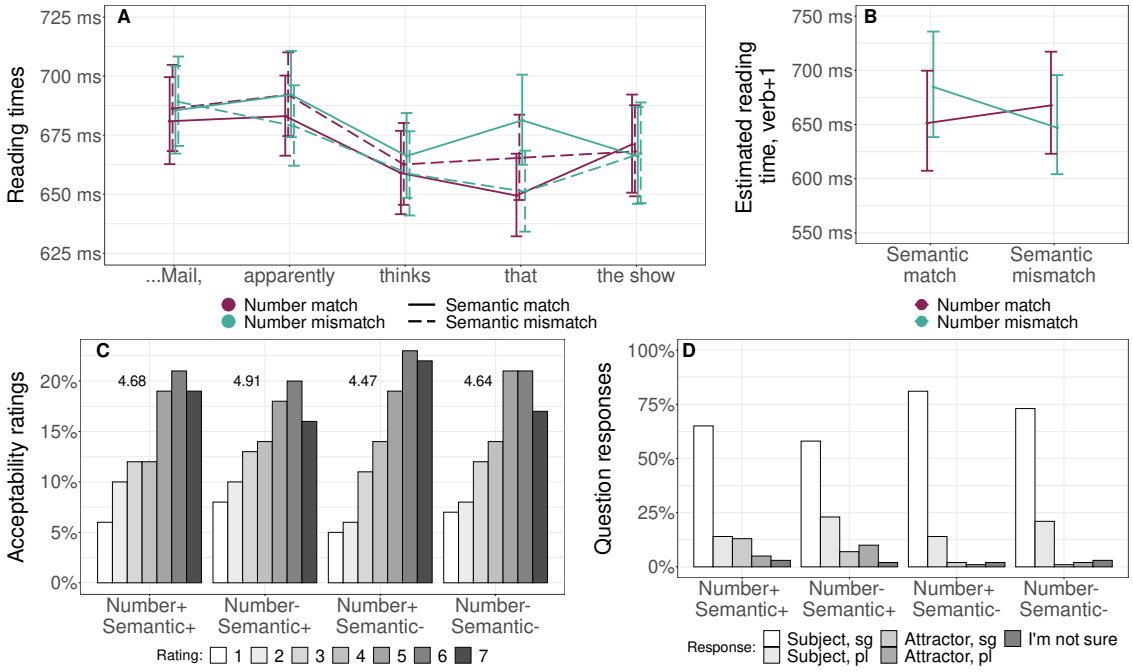

Figure 5. Results of Experiment 2. Panel A: reading times (geometric means) across sentence regions and 95% confidence intervals. Panel B: Estimated reading times at the verb+1 with 95% credible intervals. Panel C: acceptability ratings across conditions. Numbers represent mean rating in each condition. Panel D: proportions of question responses across conditions. In panels C and D, the x-axis encodes experimental conditions: *Number+* stands for number match, *Number−* for number mismatch; similarly, *Semantic+* stands for semantic match, and *Semantic−* for semantic mismatch.

$P(\beta > 0) = 0.58$; interaction: $\hat{\beta} = 5.4\,\text{ms}$, 95%-CrI: $[-16, 26]\,\text{ms}$, $P(\beta > 0) = 0.71$). On the region following the verb, there was no effect of number match ($\hat{\beta} = 6.3\,\text{ms}$, 95%-CrI: $[-14, 27]\,\text{ms}$, $P(\beta > 0) = 0.73$) or semantic match ($\hat{\beta} = 10\,\text{ms}$, 95%-CrI: $[-7.9, 28]\,\text{ms}$, $P(\beta > 0) = 0.87$). However, there was an interaction ($\hat{\beta} = 27\,\text{ms}$, 95%-CrI: $[8.2, 46]\,\text{ms}$, $P(\beta > 0) > 0.99$). Nested comparisons showed that in semantic match conditions, the number mismatch condition (b) was read more slowly than the number match condition (a) ($\hat{\beta} = 34\,\text{ms}$, 95%-CrI: $[5, 62]\,\text{ms}$, $P(\beta > 0) = 0.99$). In semantic match conditions, there was no reliable evidence for a difference ($\hat{\beta} = -21\,\text{ms}$, 95%-CrI: $[-48, 7]\,\text{ms}$, $P(\beta < 0) = 0.93$). If anything, the data suggests a (non-reliable) number mismatch effect in the opposite direction (see Figure 5B).

Acceptability ratings. The semantic match conditions had lower acceptability ratings ($\hat{\beta} = -4.5\%$, 95%-CrI: $[-9, -0.02]\%$, $P(\beta < 0) = 0.98$). Number mismatch

condition tended to receive lower ratings as well ($\hat{\beta} = -5.5\%$, 95%-CrI: $[-11, 0.35]\%$, $P(\beta < 0) = 0.97$). There was no evidence for an interaction ($\hat{\beta} = 0.66\%$, 95%-CrI: $[-2.5, 4]\%$, $P(\beta > 0) = 0.67$).

Question response accuracies. Both number mismatch ($\hat{\beta} = -8.9\%$, 95%-CrI: $[-14, -3.7]\%$, $P(\beta < 0) > 0.99$) and semantic match conditions ($\hat{\beta} = -15\%$, 95%-CrI: $[-20, -9.2]\%$, $P(\beta < 0) > 0.99$) had lower accuracies. There was no evidence for an interaction ($\hat{\beta} = 1.2\%$, 95%-CrI: $[-2.2, 4.7]\%$, $P(\beta > 0) = 0.76$).

Discussion

The aim of Experiment 2 was to eliminate the plural complexity effect on the pre-critical region in order to test the hypothesis that agreement attraction and inhibitory interference effects mask each other on average.

Long-lasting plural complexity effects were successfully eliminated. The slowdown was detected only on the plural distractor noun itself but not on any of the following words which belonged to the parenthetical (in line with the claim by Dillon et al. (2017) that parenthetical phrases are processed independently of their embedding structures). However, the masking hypothesis was again not supported. Although there was an interaction between number and semantic match/mismatch conditions, it did not fully match the predicted pattern. A slowdown due to plural distractor noun compatible with agreement attraction appeared in the semantic match conditions but not in the semantic mismatch conditions. To remind the reader, we expected attraction to be greater in the semantic match conditions but to be present in semantic mismatch conditions, as well. More importantly, the masking hypothesis clearly predicts an inhibitory semantic interference effect which was not observed.

Overall, support for similarity-based interference effects remains weak. Neither semantic nor number interference was found in reading times; comprehension question responses were compatible with semantic interference but alternative explanations could not be ruled out (see discussion of Experiment 1).

The only result unambiguously supporting semantic interference was lower acceptability ratings in the semantic match conditions.

Experiment 3

Experiment 3 uses an object relative clause construction where the distractor noun is further away from the verb than the actual subject (see (7)). The motivation for using this construction is that it introduces more material between the distractor noun and the verb and therefore, mitigates the plural complexity spill-over found in Experiment 1.

The hypotheses, number of participants, and analyses planned for Experiment 3 were pre-registered on OSF together with Experiment 2 (doi:10.17605/OSF.IO/VM5BW; on the left side menu, go to Wiki → Component Wiki pages). Procedure and analysis were the same as in Experiment 1, except for the differences in the number of participants and experimental materials described below.

Participants. Participant recruitment and exclusion procedure was the same as in Experiments 1 and 2. Participation in Experiment 3 was open for those who took part in the previous experiments as the materials were sufficiently different and the experiments were separated by at least a week. Responses from 4,914 participants were received. After applying exclusion criteria, 3,559 participants were left.

Materials. The 16 items from Experiment 1 were restructured to form sentences with object relative clauses, where the distractor noun is the head of the main clause, see Example (7)[7]:

(7) a. The singer that the actor openly admires apparently ...

b. The singers that the actor openly admires apparently ...

[7] Seven of the original object-relative items did not fully conform to our desired design. We corrected them and collected new data for these items at a later time. Time-of-day and/or day-of-the-week effects introduced by this, if any, would be absorbed by the by-item random effects and not affect our inferences.

c. The play that the actor openly admires apparently ...

d. The plays that the actor openly admires apparently ...

... received broad international recognition.

Within each sentence, the distractor noun was followed by an object relative clause containing the subject, an adverb, and a verb with correct (singular) number marking. The verb was followed by a region that did not differ across conditions and did not indicate the number of the subject noun.

The comprehension questions from Experiment 1 were modified to match the new sentences, as in Example (8):

(8) Who felt admiration? — Actor/Actors/Play/Plays/I'm not sure. Or, in semantic match conditions: Actor/Actors/Singer/Singers/I'm not sure.

Results

Summaries of reading times, acceptability ratings, and question response accuracies are presented in Figure 6. Mean question response accuracy in Experiment 3 was lower than in previous experiments (57%), which was likely due to the added complexity of processing object-relative clauses (Gibson, 2000; Gordon et al., 2001). Still, participants performed well above chance (with five response options, chance performance would be at 20% accuracy). In addition, if participants were guessing, the proportion of "I'm not sure" responses would have been much higher than the observed 5%.

Reading times. In the two regions preceding the verb, we found no main effect of number match, so we proceeded to the pre-registered analysis. At the verb, there was a pronounced slowdown in conditions with plural distractor nouns ($\hat{\beta} = 59\,\text{ms}$, 95%-CrI: $[14, 106]\,\text{ms}$, $P(\beta > 0) > 0.99$). Semantic match did not affect reading times ($\hat{\beta} = 14\,\text{ms}$, 95%-CrI: $[-15, 44]\,\text{ms}$, $P(\beta > 0) = 0.82$). There was also no evidence for an interaction ($\hat{\beta} = 0.41\,\text{ms}$, 95%-CrI: $[-31, 33]\,\text{ms}$, $P(\beta > 0) = 0.51$).

Acceptability ratings. Number mismatch conditions received substantially

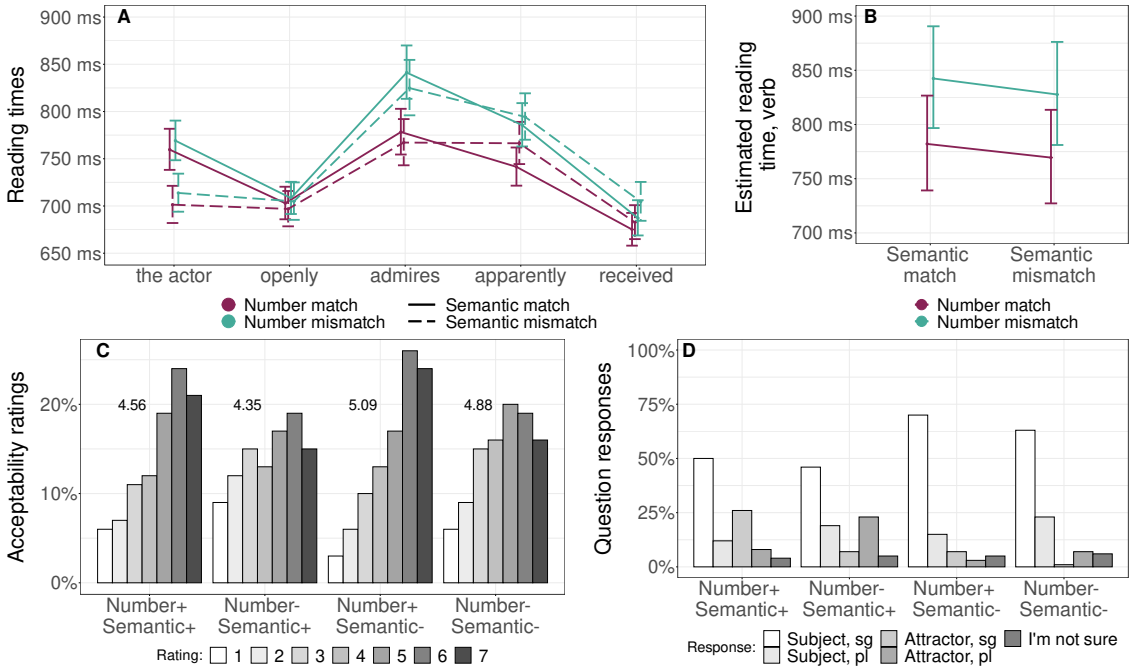

Figure 6. Results of Experiment 3. Panel A: reading times (geometric means) across sentence regions and 95% confidence intervals. Panel B: Estimated reading times at the verb with 95% credible intervals. Panel C: acceptability ratings across conditions. Numbers represent mean rating in each condition. Panel D: proportions of question responses across conditions. In panels C and D, the x-axis encodes experimental conditions: *Number+* stands for number match, *Number−* for number mismatch; similarly, *Semantic+* stands for semantic match, and *Semantic−* for semantic mismatch.

lower ratings ($\hat{\beta} = -13\%$, 95%-CrI: $[-17, -8]\%$, $P(\beta < 0) > 0.99$). There was a tendency toward lower ratings for the semantic match conditions ($\hat{\beta} = -4.5\%$, 95%-CrI: $[-9.6, 0.45]\%$, $P(\beta < 0) = 0.97$). There was no evidence for an interaction ($\hat{\beta} = 0.7\%$, 95%-CrI: $[-2.5, 3.9]\%$, $P(\beta > 0) = 0.67$).

Question response accuracies. Both number mismatch ($\hat{\beta} = -6\%$, 95%-CrI: $[-9.7, -2.3]\%$, $P(\beta < 0) > 0.99$) and semantic match conditions ($\hat{\beta} = -19\%$, 95%-CrI: $[-26, -12]\%$, $P(\beta < 0) > 0.99$) had lower accuracies. There was no evidence for an interaction ($\hat{\beta} = 1.5\%$, 95%-CrI: $[-2.4, 5.4]\%$, $P(\beta > 0) = 0.79$).

Discussion

The aim of Experiment 3 was to mitigate the long-lasting plural complexity effect found in Experiment 1 and to test the masking hypothesis. The plural complexity effect was successfully eliminated, but the masking hypothesis again

received no support. Instead, once again the data favors the feature distortion accounts: The critical verb was read more slowly in the number mismatch conditions. This slowdown was not compromised either by plural complexity effect spilling over to the critical region (Experiment 1) or by a difficult-to-interpret interaction with semantic interference (Experiment 2). The slowdown was not just statistically reliable, it was also rather large (59 ms). Acceptability ratings also support the feature distortion accounts: Lower ratings in number mismatch conditions are consistent with an illusion of ungrammaticality.

Contrary to the prediction of the similarity-based interference accounts, there was again no slowdown in semantic match conditions.

We will address the implications of these findings in the general discussion. Before that, we will discuss a possible explanation for why we found consistent agreement attraction effects even though these were difficult to find in earlier studies. We also report Experiment 4 which tests this explanation.

Interim discussion

The motivation for the three experiments presented above was to test whether parsing processes postulated by several feature distortion (Bock & Eberhard, 1993; Dempsey et al., 2022; Eberhard et al., 2005; Konieczny et al., 2004) and similarity-based interference accounts (Lewis & Vasishth, 2005; McElree, 2000) might be deployed simultaneously and their effects mask each other with stimuli typically used for testing these accounts. This hypothesis was not confirmed. Across three experiments, only attraction effects compatible with feature distortion accounts were found. In Experiments 1 and 3, a main effect was detected on the critical verb. In Experiment 2, the slowdown was detected on the subsequent word (as in many previous studies, e.g., Wagers et al., 2009) and only in the semantic match conditions. Acceptability ratings also indicated an illusion of ungrammaticality: in Experiments 1 and 3, number mismatch conditions received lower ratings even though they were equally grammatical. In Experiment 2, there

was a non-reliable tendency toward lower ratings. Overall, there is converging evidence that grammatical sentences with plural distractor nouns take longer to process and are perceived as being less well-formed than their counterparts with singular distractor nouns.

Why did many previous studies fail to find these agreement attraction effects? One systematic difference between the experiments reported here and the studies reviewed in the meta-analysis is the task that participants expect to perform. Recall that our participants had to rate the acceptability of every practice sentence, then the acceptability of the experimental sentence, and only after that encountered a comprehension question. None of the reading studies included in the meta-analysis used acceptability judgments; they all employed comprehension questions.

A possible role of the task in the emergence of attraction effects was first brought up by Franck et al. (2015), who proposed that acceptability judgment taps into agreement computation and thus drives attraction effects, while self-paced reading taps into structure building and thus drives similarity-based interference. Our data supports this proposal partially: Attraction effects arise even during self-paced reading, but perhaps only as long as the goal is to rate the sentence.

If it is the acceptability judgment task that allows the agreement attraction effects to emerge, then by manipulating the expected offline task, we should be able to make the agreement attraction effects disappear.[8]

Experiment 4

The aim of Experiment 4 was to test the hypothesis that no agreement attraction effects arise when participants expect to answer comprehension questions. The only difference between Experiment 4 and Experiment 3 is that Experiment 4 presented participants with complex practice sentences followed by comprehension questions. The procedure and analysis were otherwise identical to those of

[8] Note that the *expected* task is the key factor here: the actual task following the sentence cannot influence how the sentence is read because there is only one experimental sentence in a single-trial design, and it precedes the task.

Experiment 3; for this reason, Experiment 4 was not pre-registered separately.

Participants. Participation was open, among others, for those who took part in the previous experiments, as the experiments were separated by several months. Data from 4,576 participants were collected; after applying exclusion criteria, data from 3,535 individuals was left.

Materials. Experimental items were the same as in Experiment 3, the only difference was in the practice sentences. The new practice sentences were made more complex to match the experimental sentences and allow for complex comprehension questions. Each practice sentence contained three singular animate nouns that could potentially perform the action denoted by the verb. The distractor nouns were embedded either in a subject-extracted or in an object-extracted relative clause. Instead of an acceptability judgment task, each practice sentence was followed by a comprehension question with five response options presented in random order, as in Example (9)[9]:

(9) The priest, who had privately advised the lawyer of the art dealer, is accused of withholding information.
Who was accused? — The priest/The lawyer/The art dealer/The art dealers/I'm not sure.

In contrast to Experiment 3, the experimental sentence was followed first by a comprehension question, and after that, by an acceptability judgment task (in Experiments 1 to 3, the order was reversed). The instructions reflected the change in the materials and stated: "All sentences will be followed by a comprehension question. Some sentences will also be followed by a 1 to 7 rating scale".

Analysis. To establish whether the expected offline task affects agreement attraction effects, the interaction between task and the number match/mismatch conditions needs to be tested. We therefore analyzed the pooled data from Experiments 3 and 4; Experiment 3 (with expected acceptability judgments) was

[9] All training sentences can be found in Appendix A.

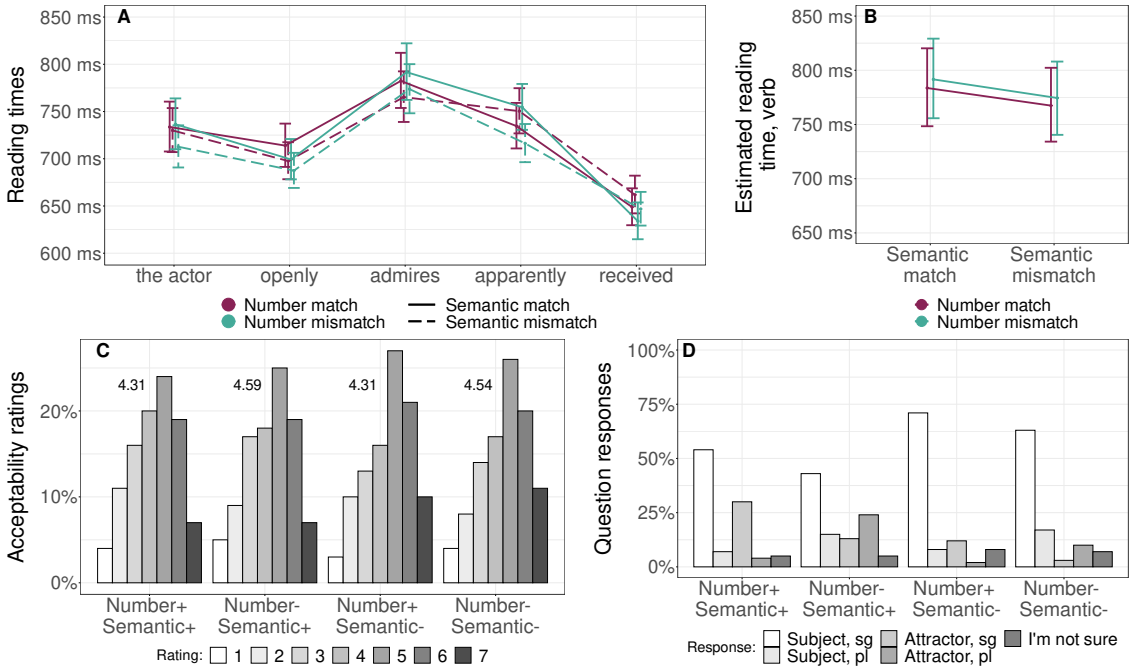

Figure 7. Results of Experiment 4. Panel A: reading times (geometric means) across sentence regions and 95% confidence intervals. Panel B: Estimated reading times at the verb with 95% credible intervals (reading times were estimated in the model including the combined data from Experiments 3 and 4; only estimates for Experiment 4 are plotted). Panel C: acceptability ratings across conditions. Numbers represent mean rating in each condition. Panel D: proportions of question responses across conditions. In panels C and D, the x-axis encodes experimental conditions: *Number+* stands for number match, *Number−* for number mismatch; similarly, *Semantic+* stands for semantic match, and *Semantic−* for semantic mismatch.

coded as 1 and Experiment 4 (with expected comprehension questions) as 0, so that the main effects of number match and semantic mismatch correspond directly to the estimates for Experiment 4. The model included all possible interactions between the experiment, number and semantic match/mismatch conditions. The random effects structure included random intercepts for items as well as by-item random slopes for all fixed effects.

Results

Summaries of reading times, acceptability ratings, and question response accuracies in Experiment 4 are presented in Figure 7. The full modeling results are presented in Appendix C; below, we report the most important findings.

Reading times. In the region preceding the verb, there was no main effect of

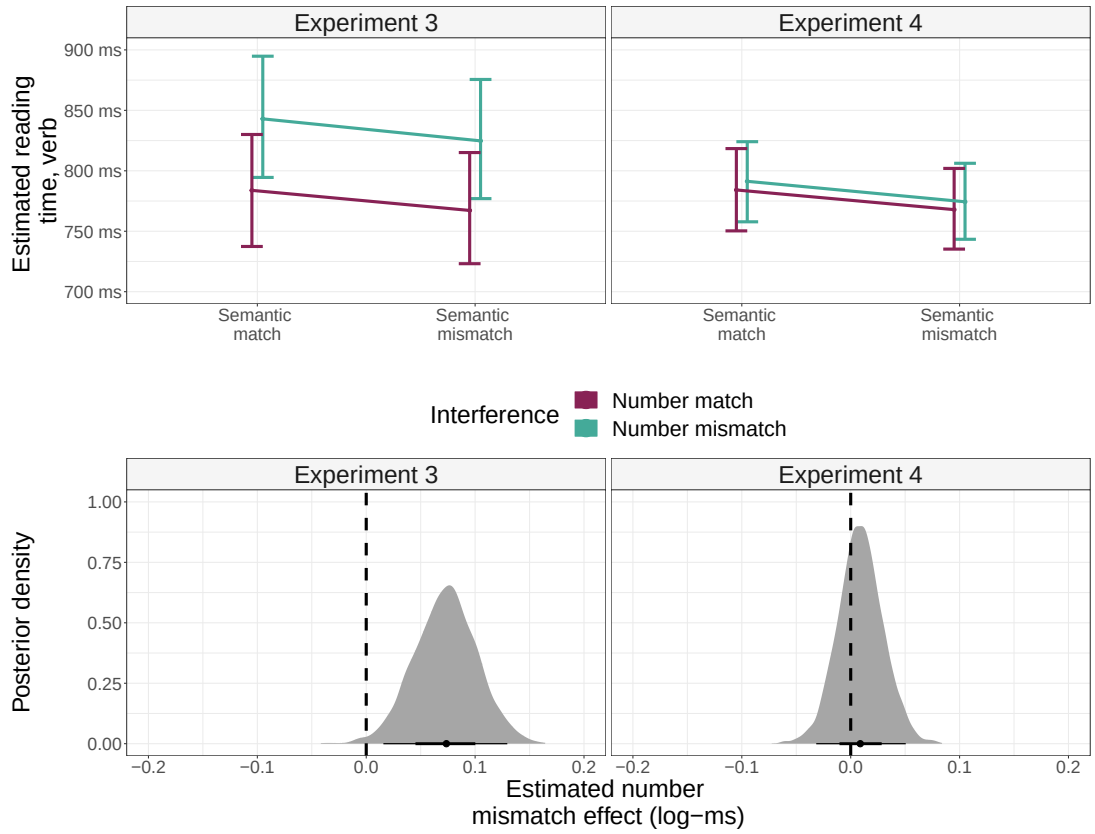

Figure 8. Attraction effect estimates in Experiment 3 vs. Experiment 4. Top row: Estimated reading times at the verb with 95% credible intervals. Bottom row: posterior density of the attraction effect estimate on the log-ms scale.

number match (or any fixed effect, see Table C1), so we proceeded to the planned analysis. As predicted, at the critical verb, there was no attraction effect in Experiment 4: $\hat{\beta} = 7\,\text{ms}$, 95%-CrI: $[-25, 39]\,\text{ms}$, $P(\beta > 0) = 0.67$ (see also Table C2 and Figure 8). There was an interaction between number mismatch and experiment ($\hat{\beta} = 50\,\text{ms}$, 95%-CrI: $[2.2, 97]\,\text{ms}$, $P(\beta > 0) = 0.98$) driven by the difference in attraction effects between the experiments (see Figure 8): In Experiment 3, the estimated slowdown due to number mismatch comprised $\hat{\beta} = 59\,\text{ms}$, 95%-CrI: $[13, 104]\,\text{ms}$, $P(\beta > 0) > 0.99$.[10] The same pattern was found in the region following the verb: in Experiment 4, there was again no reliable evidence for an attraction effect: $\hat{\beta} = -8.9\,\text{ms}$, 95%-CrI: $[-34, 15]\,\text{ms}$, $P(\beta < 0) = 0.77$. There was an

[10] The attraction effect in Experiment 3 is not estimated directly by the model. Here and elsewhere, to obtain the estimate, we combined the posterior of the attraction effect in Experiment 4 with the posteriors of the effect of experiment and the interaction between the experiment and the number match/mismatch condition (McElreath, 2020).

interaction between number mismatch and experiment ($\hat{\beta} = 42\,\text{ms}$, 95%-CrI: $[9.4, 74]\,\text{ms}$, $P(\beta > 0) > 0.99$, see also Table C3): In Experiment 3, the estimated slowdown in the number mismatch condition comprised $\hat{\beta} = 34\,\text{ms}$, 95%-CrI: $[5, 64]\,\text{ms}$, $P(\beta > 0) = 0.99$.[11]

In the region following the verb in Experiment 4, there was also an interaction between number mismatch and semantic match effects: $\hat{\beta} = 25\,\text{ms}$, 95%-CrI: $[1.5, 49]\,\text{ms}$, $P(\beta > 0) = 0.98$. However, the interaction was not the one outlined in the introduction (see Figure 3). Instead, a speedup compatible with inhibitory number interference was observed in semantic mismatch conditions: $\hat{\beta} = -34\,\text{ms}$, 95%-CrI: $[-68, 0]\,\text{ms}$, $P(\beta < 0) = 0.974$. Within semantic match conditions, there was no effect of number mismatch: $\hat{\beta} = 16\,\text{ms}$, 95%-CrI: $[-18, 51]\,\text{ms}$, $P(\beta > 0) = 0.82$. This pattern is not compatible with the predictions of either the masking hypothesis or any other sentence processing account known to us and we refrain from interpreting this effect. There was no evidence for any other fixed effect, see Table C2.

Acceptability ratings. As predicted, in Experiment 4, number mismatch did not affect ratings ($\hat{\beta} = -1\%$, 95%-CrI: $[-4.8, 2.9]\%$, $P(\beta < 0) = 0.7$). There was an interaction between number mismatch condition and experiment ($\hat{\beta} = -14\%$, 95%-CrI: $[-19, -8.1]\%$, $P(\beta < 0) > 0.99$): In Experiment 3, ratings were lower in the number mismatch condition: $\hat{\beta} = -10\%$, 95%-CrI: $[-15, -6]\%$, $P(\beta < 0) > 0.99$. In Experiment 4, ratings were lower in semantic match condition ($\hat{\beta} = -6.3\%$, 95%-CrI: $[-11, -1.4]\%$, $P(\beta < 0) > 0.99$). Ratings were overall higher in Experiment 3 than in Experiment 4 ($\hat{\beta} = 15\%$, 95%-CrI: $[2.2, 26]\%$, $P(\beta > 0) = 0.99$). There was no evidence for any other fixed effect, see Table C4.

Question response accuracies. In Experiment 4, the probability of a correct response was lower in the number mismatch ($\hat{\beta} = -9.9\%$, 95%-CrI: $[-14, -6.1]\%$, $P(\beta < 0) > 0.99$) and in semantic match conditions ($\hat{\beta} = -18\%$, 95%-CrI:

[11] Note that since more information on item-level variability is available in this pooled analysis, the estimated credible intervals got tighter, and we even detect a slowdown on the spillover region that was not detected in the separate analysis of Experiment 3.

$[-25, -12]\%$, $P(\beta < 0) > 0.99$). There was no evidence for an effect of experiment ($\hat{\beta} = 0.1\%$, 95%-CrI: $[-8.2, 8.1]\%$, $P(\beta > 0) = 0.51$) or any interaction (see Table C5).

Discussion

The goal of Experiment 4 was to test whether agreement attraction effects disappear when participants expect to read complex sentences and answer comprehension questions rather than to rate sentence acceptability. This was indeed the case. Both the slowdown in the number mismatch condition and the illusion of ungrammaticality in the acceptability ratings were absent in Experiment 4. Task effects therefore provide a plausible explanation for why the slowdown in the number mismatch condition was so rarely observed in previous studies but consistently found in Experiments 1 through 3 (evidence from Experiment 2 being more mixed). There was again no consistent support for semantic or number interference in reading times, but ratings were compatible with semantic interference.

General discussion

The present study aimed to investigate whether both feature distortion and similarity-based interference occur simultaneously in grammatical sentences, such that their opposing effects mask each other in reading times. This hypothesis received no support. Across three experiments, we found slowdowns consistent with feature distortion but not with inhibitory interference. The consistent support for feature distortion is in striking contrast to the null results prevalent in the literature. We therefore examined which feature of our experimental design might have enabled the attraction effect to occur. We evaluated one potential explanation, task adaptation, and found that when participants expected to read complex sentences and to answer complex comprehension questions, the slowdown in reading times and the illusion of ungrammaticality in acceptability judgments disappeared.

In the following sections, we address first theoretical and then practical implications of these findings.

The masking hypothesis

According to the masking hypothesis, if feature distortion and similarity-based interference operate simultaneously, either just inhibitory semantic interference or both semantic interference and agreement attraction as well as their interaction should be observed in the tested configuration. The motivation for this hypothesis was that both effects are strongly expected based on theory, but neither could be reliably demonstrated in earlier studies. In addition, previous computational simulations based on 17 datasets were compatible with this hypothesis (Yadav et al., 2023). However, the hypothesis received no support from the present data. Although an interaction between semantic and number match conditions was observed in Experiments 2 and 4 (in both cases at the postcritical region), these interaction effects did not follow the predicted pattern and also differed from each other. In sum, the masking hypothesis was not supported in any way because the consistent prediction of the combined account, inhibitory semantic interference, was not supported.

Similarity-based interference

In four experiments, we found no evidence for inhibitory number interference in reading times. If number interference and agreement attraction effects mask each other, we should have observed similar reading times in conditions with plural and singular distractor nouns. However, across four experiments, such a pattern was found only in Experiment 4 with expected comprehension questions. Possible explanations for the lack of number interference in Experiments 1 through 3 are that the number interference effect may be much smaller than the slowdown due to attraction (which corresponds to greater mismatch penalty) and/or task-dependent. Although our failure to observe interference from number may not directly contradict similarity-based interference, the occurrence of agreement attraction,

which is the opposite effect, poses a challenge to the theory (Jäger et al., 2017; Lewis & Vasishth, 2005; Yadav et al., 2023).

Inhibitory semantic interference, also predicted by the similarity-based interference accounts, was likewise not observed in any of the four experiments reported here. The lack of the predicted slowdown is consistent with several recent studies failing to detect the effect (Cunnings & Sturt, 2018; Fujita & Cunnings, 2022; Mertzen et al., 2020). With roughly 3900 participants per experiment, we should have had ~80% power to detect a 13 ms effect (a mean estimate for semantic interference obtained by Jäger et al., 2017). The chance of not finding the effect in four experiments even when it is real should therefore be $0.2^4 = 0.0016$ or 0.16%. It is therefore fairly unlikely that the absence of evidence for similarity-based interference was a false negative due to insufficient power. Although acceptability ratings and question response accuracies are broadly compatible with semantic interference, these measures might reflect not the structure building during online processing but rather post-hoc interpretative processes (Bader & Meng, 2018; Cutter et al., 2022; Dempsey et al., 2022; Meng & Bader, 2021).

To summarize, overall support for the predicted inhibitory semantic and number interference across four experiments is rather weak, consistent with recent failures to detect inhibitory interference in well-formed sentences (Avetisyan et al., 2020; Brehm et al., 2019; Cunnings & Sturt, 2018; Jäger et al., 2017; Jäger et al., 2020; Mertzen et al., 2020; Smith et al., 2021). At present, it is unclear why interference effects were not observed. They may be too small to detect (smaller than the 13 ms assumed in the power analysis), or they may be task-dependent, or restricted to the processing of ungrammatical sentences (Lago et al., 2015; Wagers et al., 2009).

Agreement attraction

Slowdowns in processing grammatical sentences with plural distractor nouns were found across three experiments, both when the distractor noun was a part of

the subject noun phrase (Experiments 1 and 2, with more mixed evidence in Experiment 2) and when it was not (Experiment 3). Attraction effects in grammatical sentences are compatible with feature distortion accounts: the feature percolation account by Bock and Eberhard (1993), the marking and morphing account by Eberhard et al. (2005, which better explains attraction effects in object relative clauses), and the attraction mechanism recently proposed by Dempsey et al. (2022). In the latter account, the presence of a particular feature, in this case, *plural*, is remembered, but the mapping between the feature and the noun decays quickly. This idea may be compatible with the lack of attraction effects when participants expected to answer comprehension questions. Speculatively, participants might have aimed to encode the sentence more accurately, which allowed the mapping between the plural feature and its host to remain active longer.

While the precise mechanisms driving attraction effects remain unclear, the important finding of the present study is that agreement attraction in grammatical sentences depends on the task. It is present when participants expect to judge sentence acceptability and disappears when participants expect to read complex sentences and answer comprehension questions. Although the task-dependency of agreement attraction does not identify the precise mechanism driving the effect, it helps to narrow down the search space of possible mechanisms. In the following section, we speculate about one such mechanism.

The role of the task

Perhaps the most surprising outcome of the present study is that changing as little as three practice sentences and the respective comprehension probes can make the effect under investigation appear or disappear. This finding aligns with studies showing that the expected task can affect underspecification of syntactically ambiguous sentences (Logačev & Vasishth, 2016; Swets et al., 2008), ambiguous pronouns (Stewart et al., 2007), and quantifier scope ambiguities (Dwivedi, 2013). More broadly, it also fits into the line of research demonstrating that the

experimental setting—the amount, distribution, and composition of fillers intermixed with experimental materials—can qualitatively change the perception of experimental materials and therefore affect the outcomes of an experiment (Arehalli & Wittenberg, 2021; Hammerly et al., 2019).

But how exactly might reading differ when the goal of the reader is rating vs. answering comprehension questions? To answer this question, we turn to computational simulations by Yadav et al. (2023). Although Yadav et al. favor the feature percolation-plus-retrieval account (equivalent to the masking hypothesis) on the grounds of its parsimony, the retrieval part of the combined account is not supported by our data. Importantly, another combined account provided an equally good fit to the data sets evaluated by Yadav et al.: namely, the combination of the feature distortion and a grammaticality bias. Grammaticality bias refers to participants' strong expectation of seeing only well-formed sentences. Hammerly et al. (2019) showed that when grammaticality bias is decreased and participants expect to encounter a lot of ungrammatical sentences, an illusion of ungrammaticality arises in binary grammaticality judgments.

Below, we sketch a proposal of how changes in grammaticality bias can explain our findings. The main idea is that the task itself can influence participants' expectations of sentence well-formedness: The acceptability rating but not the question answering task prepares participants for encountering ill-formed sentences.[12] As a consequence, participants' grammaticality bias in the sentence acceptability rating task decreases. Moreover, in Experiments 1 through 3, participants not only expected to rate sentences, they also saw an ungrammatical training sentence right before encountering the experimental sentence. Recent exposure to an ungrammatical stimulus should additionally decrease the grammaticality bias, driving the agreement attraction effect in reading times. In contrast, Experiment 4 is free from either of those influences: Comprehension questions do not bias expectation of grammaticality, and the training sentences for

[12] We are greateful to Christopher Hammerly for proposing this explanation.

Experiment 4 were all grammatical.

Our experimental data fits the grammaticality bias explanation well, but only an independent verification can confirm the causal relationship. One potential test would be to replicate Experiments 1 and 3 reported by Hammerly et al. (2019) using self-paced reading instead of rapid serial visual presentation.

Another consideration that needs to be taken into account is that the experiments differ not only in the expected task but also in expected sentence difficulty: Recall that training sentences were relatively simple in Experiments 1 through 3 vs. syntactically complex in Experiment 4. Although it is less clear how expected sentence difficulty could specifically affect attraction effects, its role cannot be ruled out at present.

In general, the finding that different task demands can affect how people process sentences could have important implications for our understanding of parsing. A fascinating question that remains to be answered is whether other well-known phenomena are also influenced by the task demands, and how this knowledge can help us develop better theories of sentence processing.

Limitations

The present study used a novel single-trial procedure that, we argue, has a number of crucial benefits. Most importantly, single-trial minimizes adaptation to the experimental materials which is known to decrease effect sizes across many dependent measures (e.g., Arehalli & Wittenberg, 2021; Brehm et al., 2021; Demberg & Sayeed, 2016; Fine et al., 2013; Ness & Meltzer-Asscher, 2021; Pregla et al., 2021). As a result, effects can be expected to be larger with single-trial designs, which, everything else being equal, implies higher statistical power and more robust conclusions (see Laurinavichyute & von der Malsburg, 2022, for a power analysis suggesting 99% power for agreement attraction effects in a forced-choice completion study). Furthermore, results from a single-trial design may be more representative of everyday language use compared to those obtained with

analogous repeated-measures designs in which participants see the same structures over and over again. A single-trial procedure also necessarily taps into a much larger and therefore more diverse pool of participants, which means that findings can be expected to generalize more broadly, e.g., beyond psychology undergraduates (Yarkoni, 2022).[13]

However, single-trial procedures may have downsides, as well. In the present case, the training sentences were fixed and may have had a substantial impact on how participants processed the subsequent experimental sentence. In an analogous repeated-measures design, the sentences preceding the experimental sentences vary and such predecessor effects may therefore wash out. This, of course, comes at the price of having to repeat the same structures many times which may affect ecological validity. Nonetheless, the peculiarities and implications of a single-trial design must be carefully considered and there remains room for improvements of this approach. For instance, a future study may vary the training sentences to make effects in the critical sentence less dependent on them.

Conclusion

This study attempted to explain why earlier research failed to provide consistent evidence for agreement attraction effects and similarity-based interference effects that are both strongly expected based on theory. Specifically, we investigated the hypothesis that feature distortion and similarity-based interference might be at work simultaneously and their opposing effects mask each other in averaged reading times. This hypothesis was not confirmed. Across three experiments, we observed agreement attraction effects, i.e. higher processing times and lower acceptability ratings in grammatical sentences with plural distractor nouns, consistent with the predictions of feature distortion accounts. However, agreement attraction effects emerged only when participants expected to read simple sentences and rate their

[13] Von der Malsburg et al. 2020 have tested more than 25K participants from the US and UK on Mechanical Turk and Prolific and found in both cases that demographics (education, age, gender, political alignment) were close to those found in the respective general populations.

acceptability (Experiments 1–3); when participants expected to read more complex sentences and answer comprehension questions (Experiment 4), agreement attraction effects disappeared. These findings suggest that at least some phenomena in sentence processing may be more labile than previously thought and may depend on seemingly small details of the experimental setting such as the precise task that participants expect to perform. If true, conclusions based on effects observed in highly constrained and repetitive experiments may be less generalizable to everyday language processing than is usually assumed.

Acknowledgements. We would like to thank Kate M. Stone and Garrett Smith for their help with creating experimental items, and to the audience of CUNY 2018, CUNY 2019, CUNY 2021, and LISP for helpful feedback.

Funding statements. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 317633480, SFB 1287. Data collection for Experiment 1, as well as 1960 participants in each of Experiments 2 and 3 had been funded by the structural division Cognitive Science (SBKW), University of Potsdam.

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American psychologist*, *51*(4), 355.

Arehalli, S., & Wittenberg, E. (2021). Experimental filler design influences error correction rates in a word restoration paradigm. *Linguistics Vanguard*, *7*(1), 20200052.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, 104087.

Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1286.

Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, *8*(1), 57–99.

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze Made Easy: Better and easier measures of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082.

Brehm, L., Jackson, C. N., & Miller, K. L. (2019). Speaker-specific processing of anomalous utterances. *Quarterly Journal of Experimental Psychology*, *72*(4), 764–778.

Brehm, L., Jackson, C. N., & Miller, K. L. (2021). Probabilistic online processing of sentence anomalies. *Language, Cognition and Neuroscience*, 1–25.

Bürkner, P.-C., et al. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Cunnings, I., & Sturt, P. (2018). Retrieval interference and semantic interpretation. *Journal of Memory and Language*, *102*, 16–27.

Cutter, M. G., Paterson, K. B., & Filik, R. (2022). Online representations of non-canonical sentences are more than good-enough. *Quarterly Journal of Experimental Psychology*, *75*(1), 30–42.

Demberg, V., & Sayeed, A. (2016). The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PloS one*, *11*(1), e0146194.

Dempsey, J., Christianson, K., & Tanner, D. (2022). Misretrieval but not misrepresentation: A feature misbinding account of post-interpretive effects in number attraction. *Quarterly Journal of Experimental Psychology*, *75*(9), 1727–1745.

Dillon, B., Clifton, C., Sloggett, S., & Frazier, L. (2017). Appositives and their aftermath: Interference depends on at-issue vs. not-at-issue status. *Journal of Memory and Language*, *96*, 93–109.

Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*, *8*(11), e81461.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *43*(12), e12800.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one*, *8*(10), e77661.

Franck, J., Colonna, S., & Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in psychology*, *6*, 349.

Fujita, H., & Cunnings, I. (2022). Interference and filler-gap dependency formation in native and non-native language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(5), 702.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, *2000*, 95–126.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology: learning, memory, and cognition*, *27*(6), 1411.

Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive psychology*, *110*, 70–104.

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., & Vasishth, S. (2015). Teasing apart retrieval and encoding interference in the processing of anaphors. *Frontiers in psychology*, *6*, 506.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339.

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*, 104063.

Kay, M. (2019). *tidybayes: Tidy data and geoms for Bayesian models* [R package version 1.1.0].

Konieczny, L., Schimke, S., & Hemforth, B. (2004). An activation-based model of agreement errors in production and comprehension. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26).

Lago, S., Acuña Fariña, C., & Meseguer, E. (2021). The reading signatures of agreement attraction. *Open Mind*, *5*, 132–153.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

Laurinavichyute, A. (2021). *Similarity-based interference and faulty encoding accounts of sentence processing* (doctoralthesis). Universität Potsdam.

Laurinavichyute, A., Jäger, L. A., Akinina, Y., Roß, J., & Dragoy, O. (2017). Retrieval and encoding interference: Cross-linguistic evidence from anaphor processing. *Frontiers in psychology*, *8*, 965.

Laurinavichyute, A., & von der Malsburg, T. (2022). Semantic attraction in sentence comprehension. *Cognitive Science*, *46*(2), e13086.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, *29*(3), 1–45.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, *10*(10), 447–454.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Logačev, P., & Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*(2), 266–298.

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

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of psycholinguistic research*, *29*(2), 111–123.

Meng, M., & Bader, M. (2021). Does comprehension (sometimes) go wrong for noncanonical sentences? *Quarterly Journal of Experimental Psychology*, *74*(1), 1–28.

Mertzen, D., Laurinavichyute, A., Dillon, B. W., Engbert, R., & Vasishth, S. (2020). *Crosslinguistic evidence against interference from extra-sentential distractors* [submitted].

Ness, T., & Meltzer-Asscher, A. (2021). Rational adaptation in lexical prediction: The influence of prediction strength. *Frontiers in Psychology*, *12*.

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive science*, *42*, 1075–1100.

Nicol, J., Forster, K., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, *36*(4), 569–587.

Paape, D., Avetisyan, S., Lago, S., & Vasishth, S. (2021). Modeling misretrieval and feature substitution in agreement attraction: A computational evaluation. *Cognitive Science*, *45*(8), e13019.

Parker, D., & An, A. (2018). Not all phrases are equally attractive: Experimental evidence for selective agreement attraction effects. *Frontiers in psychology*, *9*, 1566.

Patson, N. D., & Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *The Quarterly Journal of Experimental Psychology*, *69*(5), 950–971.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and language*, *41*(3), 427–456.

Pregla, D., Lissón, P., Vasishth, S., Burchert, F., & Stadie, N. (2021). Variability in sentence comprehension in aphasia in German. *Brain and Language*, *222*, 105008.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113.

Schlueter, Z., Parker, D., & Lau, E. F. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in psychology*, *10*, 1002.

Slioussar, N. (2018). Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language*, *101*, 51–63.

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, *124*, 101356.

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of memory and language*, *60*(2), 308–327.

Stewart, A. J., Holler, J., & Kidd, E. (2007). Shallow processing of ambiguous pronouns: Evidence for delay. *Quarterly Journal of Experimental Psychology*, *60*(12), 1680–1696.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216.

Tendeiro, J. N., & Kiers, H. A. L. (2022). With bayesian estimation one can get all that Bayes factors offer, and more. *Psychonomic Bulletin & Review*.

Thornton, R., & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language*, *48*(4), 740–759.

Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in psychology*, *6*, 347.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 407.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of memory and language*, *65*(3), 247–263.

Vasishth, S. (2006). On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. *Proceedings of the International Conference on Linguistic Evidence*, 96–100.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of phonetics*, *71*, 147–161.

Veríssimo, J. (2021). Analysis of rating scales: A pervasive problem in bilingualism research and a solution with Bayesian ordinal models. *Bilingualism: Language and Cognition*.

Villata, S., & Franck, J. (2020). Similarity-based interference in agreement comprehension and production: Evidence from object agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(1), 170.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

Wagers, M. W. (2008). *The structure of memory meets memory for structure in linguistic cognition*. University of Maryland, College Park.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer.

Yadav, H., Smith, G., Reich, S., & Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, *129*, 104400.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*.

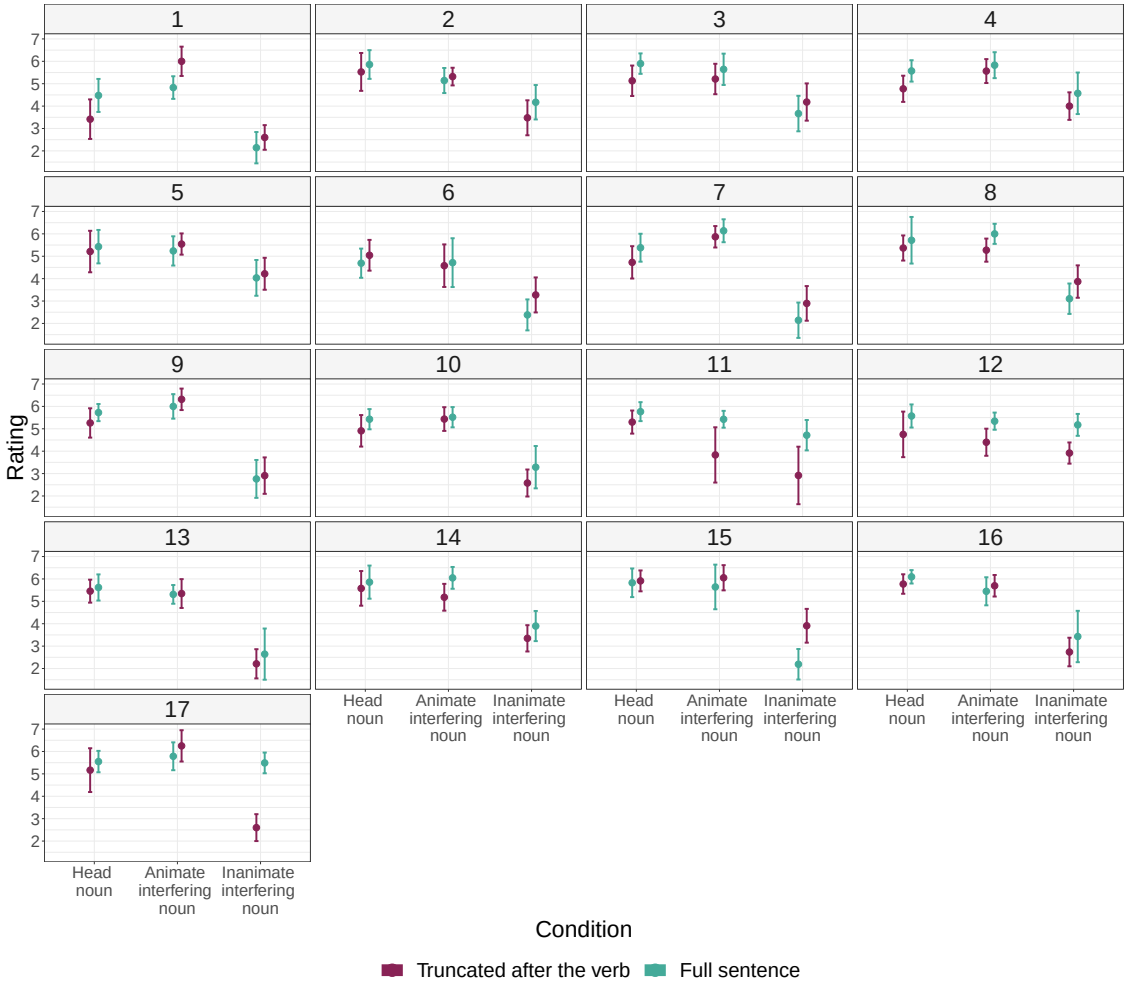

Figure A1. Mean rating for each condition across pretests and experimental items. Error bars represent 95% confidence intervals.

Appendix A

Experimental items

Item norming

Mean by-item ratings are presented on Figure A1. For statistical analysis of Likert scale ratings, we employed ordinal logistic mixed-effects models. The results of statistical analysis are presented in Table A1.

We tested 17 items while only 16 were needed for the experiment, so we decided to exclude item 11 based on the lower mean ratings and personal judgment.

Table A1
Statistical modeling of plausibility norming, the estimates are presented on the log-odds scale.

	Truncated sentences		Full sentences	
Predictor	Estimate	95%-CrI	Estimate	95%-CrI
Intercept[1]	−4.14	[−4.58, −3.70]	−3.84	[−4.30, −3.39]
Intercept[2]	−2.97	[−3.36, −2.59]	−2.72	[−3.11, −2.32]
Intercept[3]	−2.13	[−2.48, −1.77]	−2.06	[−2.43, −1.67]
Intercept[4]	−1.35	[−1.69, −1.02]	−1.05	[−1.40, −0.69]
Intercept[5]	−0.53	[−0.84, −0.20]	0.05	[−0.31, 0.39]
Intercept[6]	0.89	[0.58, 1.23]	1.58	[1.22, 1.94]
Semantic match	0.06	[−0.23, 0.35]	0.25	[−0.11, 0.57]
Semantic mismatch	−0.99	[−1.45, −0.47]	−1.21	[−1.71, −0.65]

Training sentences

Training sentences used in Experiments 1, 2, and 3

(10) This is a practice sentence to get you used to reading sentences like this.
Was that a good sentence? — Likert rating scale from 1 to 7.

(11) The donation the veteran made was doubled by an unknown philanthropist.
Was that a good sentence? — Likert rating scale from 1 to 7.

(12) The books was borrowed by the local expert.
Was that a good sentence? — Likert rating scale from 1 to 7.

Training sentences used in Experiment 4

(13) The priest, who had privately advised the lawyer of the art dealer, is accused of withholding information.
Who was accused? — The priest/The lawyer/The art dealer/The art dealers/I'm not sure.

(14) The personal assistant, who the bodyguard of the delegate does not trust, attracts great public attention.
Who attracted public attention? — The personal assistant/The bodyguard/The delegate/The bodyguards/I'm not sure.

(15) The philanthropist, who had greeted the secretary of the director, later participated in the fundraising committee.
Who took part in the committee? — The philanthropist/The secretary/The director/The secretaries/I'm not sure.

Appendix B

Statistical modeling

The priors used for the models and their interpretation can be seen in Table B1. For the parameters not specified in Table B1, default brms priors were used. The models were run using four chains, with the default 2000 iterations per chain unless increasing the number of iterations was recommended. In such cases, each chain ran for 6000 iterations. For all models, the first half of iterations were discarded as warm-up iterations. For all models, convergence diagnostics indicated no convergence issues, i.e. $\hat{R}$ values were close to 1.

Table B1
Prior specifications for the models and their explanation. Reading times are modeled on the log-normal scale, acceptability ratings and accuracies on the log-odds scale.

Parameter	Prior	Interpretation
Meta-analysis		
Intercept	$\mathcal{N}(\mu = 0, \sigma = 100)$	The overall effect likely lies between −195 ms and 195 ms.
sd(Intercept)	$\mathcal{N}(\mu = 0, \sigma = 100)$	By-item deviation of any effect from the overall estimate likely lies between 0 ms and 225 ms.
Reading times		
Intercept	$\mathcal{N}(\mu = 6.5, \sigma = 0.7)$	RTs likely lie between 170 ms and 2700 ms.
β	$\mathcal{N}(\mu = 0, \sigma = 0.2)$	The estimated effects likely lie between −880 ms and 880 ms.
sd(β)	$\mathcal{N}(\mu = 0, \sigma = 0.5)$	By-item deviation of any effect from the overall estimate likely lies between −690 ms and 1805 ms.
σ	$\mathcal{N}(\mu = 0, \sigma = 1)$	Residual likely lies between 15 ms and 9400 ms.
Accuracies		
Intercept	$\mathcal{N}(\mu = 0, \sigma = 1.5)$	Accuracy likely lies between 5% and 95%.
β	$\mathcal{N}(\mu = 0, \sigma = 0.3)$	The estimated effects likely lie between −24% and 24%.
sd(β)	$\mathcal{N}(\mu = 0, \sigma = 0.5)$	By-item deviation of any effect from the overall estimate likely lies between −20% and 20%.
Acceptability ratings		
Intercept	$\mathcal{N}(\mu = 0, \sigma = 1.5)$	Each rating is equally possible, and the most likely probability of each rating is 14%.
β	$\mathcal{N}(\mu = 0, \sigma = 0.3)$	Independent variables can shift rating probabilities by up to 25%.
sd(β)	$\mathcal{N}(\mu = 0, \sigma = 0.5)$	By-item deviation of any effect from the overall estimate likely lies between −32% and 32%.

Appendix C

Joint analysis of Experiment 3 and Experiment 4

Table C1
Joint analysis of Experiment 3 and Experiment 4. Statistical modeling of reading times at the precritical region.

Predictor	Estimate (log-ms)	95%-CrI	$P(\beta > 0)$
Intercept	6.55	$[6.50; 6.60]$	>0.999
Number mismatch in Exp. 4	−0.01	$[-0.02; 0.01]$	0.13
Semantic match in Exp. 4	0.01	$[-0.00; 0.02]$	0.93
Experiment 3	0.01	$[-0.04; 0.06]$	0.62
Number mismatch × Semantic match in Exp. 4	−0.00	$[-0.01; 0.01]$	0.47
Number mismatch × Experiment 3	0.01	$[-0.01; 0.03]$	0.87
Semantic match × Experiment 3	−0.02	$[-0.13; 0.09]$	0.25
Number mismatch × Semantic match × Exp. 3	−0.00	$[-0.02; 0.02]$	0.32

Table C2
Joint analysis of Experiment 3 and Experiment 4. Statistical modeling of reading times at the critical region.

Predictor	Estimate (log-ms)	95%-CrI	$P(\beta > 0)$
Intercept	6.66	$[6.63; 6.68]$	>0.999
Number mismatch in Exp. 4	0.00	$[-0.02; 0.03]$	0.67
Semantic match in Exp. 4	0.01	$[-0.01; 0.03]$	0.88
Experiment 3	0.03	$[-0.01; 0.07]$	0.93
Number mismatch × Semantic match in Exp. 4	0.00	$[-0.02; 0.02]$	0.51
Number mismatch × Experiment 3	0.03	$[0.00; 0.06]$	0.98
Semantic match × Experiment 3	−0.00	$[-0.03; 0.02]$	0.42
Number mismatch × Semantic match × Exp. 3	−0.00	$[-0.03; 0.03]$	0.50

Table C3
Joint analysis of Experiment 3 and Experiment 4. Statistical modeling of reading times at the postcritical region.

Predictor	Estimate (log-ms)	95%-CrI	$P(\beta > 0)$
Intercept	6.60	$[6.55; 6.66]$	>0.999
Number mismatch in Exp. 4	−0.01	$[-0.02; 0.01]$	0.23
Semantic match in Exp. 4	0.01	$[-0.01; 0.02]$	0.82
Experiment 3	0.05	$[-0.00; 0.10]$	0.97
Number mismatch × Semantic match in Exp. 4	0.02	$[0.00; 0.03]$	0.98
Number mismatch × Experiment 3	0.03	$[0.01; 0.05]$	0.99
Semantic match × Experiment 3	−0.02	$[-0.04; 0.01]$	0.06
Number mismatch × Semantic match × Exp. 3	−0.01	$[-0.04; 0.01]$	0.12

Table C4
Statistical modeling of acceptability ratings on the data pooled from Experiments 3 and 4.

Predictor	Estimate (log-odds)	95%-CrI	$P(\beta < 0)$
Intercept[1]	−2.96	[−3.23; −2.69]	>0.99
Intercept[2]	−1.76	[−2.01; −1.50]	>0.99
Intercept[3]	−0.82	[−1.08; −0.56]	>0.99
Intercept[4]	−0.07	[−0.33; 0.18]	0.72
Intercept[5]	0.91	[0.65; 1.17]	< 0.001
Intercept[6]	2.21	[1.94; 2.48]	< 0.001
Number mismatch in Exp. 4	−0.02	[−0.10; 0.06]	0.70
Semantic match in Exp. 4	−0.13	[−0.23; −0.03]	0.99
Experiment 3	0.30	[0.04; 0.53]	0.01
Number mismatch × Semantic match in Exp. 4	0.01	[−0.04; 0.07]	0.32
Number mismatch × Experiment 3	−0.28	[−0.39; −0.16]	>0.999
Semantic match × Experiment 3	0.02	[−0.10; 0.13]	0.33
Number mismatch × Semantic match × Exp. 3	−0.00	[−0.09; 0.09]	0.50

Table C5
Statistical modeling of question response accuracies on the data pooled from Experiments 3 and 4.

Predictor	Estimate (log-odds)	95%-CrI	$P(\beta < 0)$
Intercept	0.35	[0.08; 0.61]	0.0065
Number mismatch in Exp. 4	−0.20	[−0.28; −0.12]	> 0.999
Semantic match in Exp. 4	−0.39	[−0.52; −0.24]	> 0.999
Experiment 3	0.00	[−0.09; 0.09]	0.49
Number mismatch × Semantic match in Exp. 4	0.00	[−0.16; 0.16]	0.77
Number mismatch × Experiment 3	0.08	[−0.02; 0.18]	0.06
Semantic match × Experiment 3	−0.02	[−0.13; 0.09]	0.64
Number mismatch × Semantic match × Exp. 3	0.06	[−0.04; 0.16]	0.13