COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13086

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13086

# Semantic Attraction in Sentence Comprehension

Anna Laurinavichyute,[a,b] Titus von der Malsburg[c,d]

[a]*Department of Linguistics, University of Potsdam*
[b]*Center for Language and Brain, HSE University*
[c]*Institute of Linguistics, University of Stuttgart*
[d]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

Received 14 January 2021; received in revised form 9 December 2021; accepted 14 December 2021

**Abstract**

Agreement attraction is a cross-linguistic phenomenon where a verb occasionally agrees not with its subject, as required by grammar, but instead with an unrelated noun ("The key to the cabinets were…"). Despite the clear violation of grammatical rules, comprehenders often rate these sentences as acceptable. Contenders for explaining agreement attraction fall into two broad classes: Morphosyntactic accounts specifically designed to explain agreement attraction, and more general sentence processing models, such as the Lewis and Vasishth model, which explain attraction as a consequence of how linguistic structure is stored and accessed in content-addressable memory. In the present research, we disambiguate between these two classes by testing a surprising prediction made by the Lewis and Vasishth model but not by the morphosyntactic accounts, namely, that attraction should not be limited to morphosyntax, but that semantic features of unrelated nouns equally induce attraction. A recent study by Cunnings and Sturt provided initial evidence that this may be the case. Here, we report three single-trial experiments in English that compared semantic and agreement attraction and tested whether and how the two interact. All three experiments showed strong semantically induced attraction effects closely mirroring agreement attraction effects. We complement these results with computational simulations which confirmed that the Lewis and Vasishth model can faithfully reproduce the observed results. In sum, our findings suggest that attraction is a more general phenomenon than is

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 317633480, SFB 1287. A.L. was also supported by the Center for Language and Brain, NRU Higher School of Economics, RF Government Grant 14.641.31.0004.

Correspondence should be sent to Anna Laurinavichyute, Department of Linguistics, University of Potsdam, Haus 14, Karl-Liebknecht-Strabe 24-25,14476 Potsdam, Germany. E-mail: anna.laurinavichyute@uni-potsdam.de

commonly believed, and therefore favor more general sentence processing models, such as the Lewis and Vasishth model.

*Keywords:* Agreement attraction; Computational modeling; Sentence processing; Similarity-based interference; Semantic attraction

---

One way to understand how the human language processing system operates is to study the errors language users make and the circumstances that affect these errors. One particularly well-studied type of error is called agreement attraction (Bock & Miller, 1991; Kimball & Aissen, 1971). Agreement attraction refers to erroneous agreement typically between a verb and a non-subject noun that seizes morphosyntactic control of the verb from the subject, as in:

(1) *The difference between the studies stem from …

Here, the verb agrees with "studies"—both are plural—instead of with the subject "difference," which is singular. Even though the resulting sentence is clearly ungrammatical, such sentences are regularly produced (Haskell & MacDonald, 2005) and these agreement errors often go unnoticed in comprehension (Clifton, Frazier, & Deevy, 1999; Tanner & Bulkes, 2015).

Agreement attraction has been studied intensively in language production. This research has identified various constraints on agreement attraction. For instance, agreement attraction has been found more reliably when the subject is singular, as in (1), than when it is plural (referred to as *singular–plural* asymmetry, see, for example, Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991; Deutsch & Dank, 2011; Eberhard, 1997, but see Franck, Vigliocco, & Nicol, 2002, for a counterexample). While the position of the attractor seems to have some impact on the strength of agreement attraction (e.g., Franck, Lassi, Frauenfelder, & Rizzi, 2006; Franck et al., 2002), there is currently little evidence suggesting that syntactic constraints can completely prevent a noun from interfering with the subject–verb dependency (but see Franck, Soare, Frauenfelder, & Rizzi, 2010, who report some evidence for immunity to agreement attraction in complement clauses). Agreement attraction has also been demonstrated in a variety of languages other than English, and there is some evidence that languages with richer morphosyntax, for example, Russian and Spanish, may be more robust to agreement attraction (Foote & Bock, 2012; Lorimor, Bock, Zalkind, Sheyman, & Beard, 2008, but see Slioussar & Malko, 2016). Finally, it has been found that patterns of agreement attraction errors in production largely mirror the effects in comprehension, which suggests that the underlying mechanisms may be the same in production and comprehension (Pearlmutter, Garnsey, & Bock, 1999).

All models designed to explain agreement attraction in production share the assumption that attraction manifests only on the morphosyntactic level of language organization, that is, attraction is caused by mechanisms that can derail only the formation of morphosyntactic relationships in a sentence, i.e. agreement (Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001; Eberhard, Cutting, & Bock, 2005; Franck et al., 2002). According to these models, agreement attraction is a phenomenon with a rather narrow scope.

Meanwhile, more general language processing models have been used to explain agreement attraction errors in comprehension. For instance, the Lewis and Vasishth model (Engelmann, Jäger, & Vasishth, 2019; Lewis & Vasishth, 2005; Nicenboim & Vasishth, 2018) postulates that attraction errors arise from the particular way in which linguistic structure is stored in content-addressable memory. While this model has so far only been used to explain morphosyntactic attraction effects, the principles it is based on are thought to be more general. However, if we assume, as the Lewis and Vasishth model does, that agreement attraction arises from general principles of how working memory is accessed, there is no reason why attraction should be limited to the morphosyntactic level. Instead, we would expect that attraction effects should arise in other linguistic domains as well, for instance, on the level of meaning. And indeed, initial experimental evidence reported by Cunnings and Sturt (2018) is consistent with that view.

To further test whether attraction effects arise on the level of meaning and to expand on the initial evidence reported by Cunnings and Sturt, we ran three experiments in which participants had to decide whether a verb was a viable continuation for a given sentence fragment. Specifically, we tested whether language users would accept semantically mismatching verbs as viable sentence completions when there was another noun (the attractor) that satisfied the verb's demand for a semantically matching subject. More concretely, we tested (among other things) whether the singular verb form "cuts" would be accepted more often as a completion for fragments like (2-a) than for fragments like (2-b) even though "cuts" thematically fits the subject equally badly in both sentences.

(2) a. The drawer with the knife … (cuts?)
b. The drawer with the handle … (cuts?)

If attraction is limited to morphosyntax, we expect no difference in completions for (2-a) and (2-b). However, if there were more errors in (2-a) than (2-b), this would constitute evidence for attraction being a more general phenomenon than is often assumed. This finding would favor theories of sentence processing that are not isolated to narrow domains but apply across all domains of language organization. Hence, this research not only asks questions about attraction phenomena in particular, but also promises new insights into the mental representation of linguistic structure and the modularity of linguistic processes more broadly.

In the following, we will briefly review the most influential accounts of agreement attraction in production and comprehension and then outline their predictions with regard to semantic attraction errors. Then, we will report three experiments and follow up with a computational simulation examining the Lewis and Vasishth model's predictions at a more fine-grained level.

### *1.1. Production accounts of agreement attraction*

The *Feature Percolation* account was formulated to explain attraction effects in the number domain and heavily relies on the notion of markedness (Franck et al., 2002; Nicol, Forster, & Veres, 1997; Vigliocco & Nicol, 1998). Singular is considered an unmarked member of the number opposition, while plural is assumed to be marked (e.g., Bock & Eberhard, 1993;

Eberhard, 1997; Harley & Ritter, 2002). The key idea is that in sentences like (1), where the attractor noun ("studies") is part of a complex subject noun phrase ("The difference between the studies"), the plural feature of the attractor can erroneously "percolate" up the syntactic tree and override the correct number marking of the complex noun phrase. As a result, a plural verb is produced (e.g., "stem") even though the subject ("The difference") requires singular. Thus, Feature Percolation posits that the culprit is faulty encoding of the subject noun phrase, not the agreement computation itself.

The beauty of this account is its parsimony and the fact that it makes rich predictions about the circumstances under which agreement attraction can arise. For instance, Feature Percolation correctly predicts more attraction errors when the subject noun is singular than when it is plural (Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991; Deutsch & Dank, 2011; Eberhard, 1997). Another strong prediction is that agreement attraction arises only in configurations where the attractor is embedded within the subject noun phrase, such as in (1) and (3).

(3) The soldier that the officers accused …*were

However, studies have also shown agreement attraction effects in constructions where the attractor is located outside the subject–noun phrase ("The cabinets that the key …*open," Staub, 2009, 2010), in questions ("*Are the helicopter for the flights safe?," Vigliocco & Nicol, 1998), and in direct object constructions (Dutch subject–object–verb constructions: Hartsuiker, Antón-Méndez, & Van Zee, 2001; French object–subject–verb cleft constructions: Franck et al., 2006). These findings pose problems for Feature Percolation.

Further, Feature Percolation cannot explain why attraction errors increase when the subject is syntactically singular but denotes a set of items as in "The label on the bottles … *were" (Foote & Bock, 2012; Hartsuiker, Kolk, & Huinck, 1999; Vigliocco, Butterworth, & Semenza, 1995; Vigliocco, Hartsuiker, Jarema, & Kolk, 1996), or "The team with the red shirts … *were" (Humphreys & Bock, 2005; Smith, Franck, & Tabor, 2018; Solomon & Pearlmutter, 2004).

An alternative account that can explain the latter class of cases is the *Marking and Morphing* account (Bock et al., 2001; Eberhard et al., 2005). Like Feature Percolation, it assumes faulty encoding of the subject, but unlike Feature Percolation it relies on the concept of *notional number*—a semantic representation of the entity that is referred to, either as a multitude or as a single unit. Both nouns, such as "team," and noun phrases, such as "the label on the bottles," can be notionally plural while being syntactically singular. The Marking and Morphing account builds upon Feature Percolation and postulates that the subject's notional number influences the computation of number agreement over and above the morphosyntactic number match between the attractor and the verb. Essentially, the more items the subject denotes, the higher is the probability of a plural verb. Just as Feature Percolation, the account is well suited for explaining agreement attraction effects in the number domain. And like Feature Percolation, it only covers the configurations where the attractor is located within the subject noun phrase (although, unlike Feature Percolation, it could potentially be extended to cover object attraction as proposed in Eberhard et al., 2005; for recent evidence, see also Avetisyan, Lago, & Vasishth, 2020, Exp. 1).

A shortcoming of Feature Percolation and Marking and Morphing is that while both models can account for instances of gender attraction in systems with two genders (Antón-Méndez, Nicol, & Garrett, 2002; Vigliocco & Franck, 2001, 1999), it is unclear how the models could be extended to account for gender and case attraction effects in systems with more than two possible feature values (Badecker & Kuminiak, 2007; Bader & Meng, 1999; Slioussar & Malko, 2016; Slioussar, Stetsenko, & Matyushkina, 2015). In particular, notional marking is less motivated for grammatical gender, and not at all for case. In addition, the current formulation of the model is specifically tailored for binary features systems, and extending it to features with more than two values, such as case, is not straightforward.

While both Feature Percolation and Marking and Morphing were designed to explain attraction errors in production, they were also invoked to explain analogous effects in sentence comprehension (Pearlmutter et al., 1999; Wagers, Lau, & Phillips, 2009). Consider the following example sentences:

(4) a. *The key to the cells were …
b. *The key to the cell were …

In sentences with agreement errors, such as (2-a), where the intervening noun matches the verb in number, reading times were shown to be faster than in control sentences, (2-b), where the intervening noun does not match the verb in number (see also Avetisyan et al., 2020; Lago, Shalom, Sigman, Lau, & Phillips, 2015; Tucker, Idrissi, & Almeida, 2015; Villata, Tabor, & Franck, 2018, for similar results in Spanish, Eastern Armenian, Arabic, and Italian). In addition, sentences like (2-a) are more often judged as grammatical or acceptable than sentences like (2-b) (Hammerly, Staub, & Dillon, 2019; Patson & Husband, 2016; Vasishth, Jäger, & Nicenboim, 2017; Wagers et al., 2009). Feature Percolation and Marking and Morphing both explain these effects by assuming that the plural feature of the attractor in (2-a) sometimes compromises the subject's number marking, in which case the unlicensed plural verb is actually expected, and consequently, does not cause as much processing difficulty as in the control condition (2-b).

### *1.2. Comprehension theory of agreement attraction*

We will now briefly review a general model of language comprehension that can potentially explain attraction effects in comprehension, even though it was not explicitly designed for this purpose. The *Lewis and Vasishth model* (Lewis & Vasishth, 2005) is based on the content-addressable memory architecture ACT-R (Anderson, 1996). The model assumes that syntactic chunks are activated in working memory when they are encountered and later retrieved in order to build syntactic dependencies. Syntactic chunks (including words) are represented as bundles of features and are retrieved by querying a subset of these features that is relevant at the moment of retrieval (so-called retrieval cues). The model was first applied to sentences with agreement attraction errors by Wagers et al. (2009).

To understand how the Lewis and Vasishth model can explain agreement attraction, consider the grammatical sentence (5) from Wagers et al. study:

(5) The cabinets that the key opens …

When encountering the verb "opens," the parser triggers a retrieval of the subject to complete the subject–verb dependency. The verb is marked for number, and the parser therefore spreads activation to every word that has the features +SUBJECT[1] and +SINGULAR. The word with the highest activation is then retrieved and completes the dependency (if it exceeds the so-called *retrieval activation threshold*). In (5), the only word that fully matches the retrieval cues is the subject "key" and it is therefore consistently retrieved.

Now consider the sentence (4-a) that contains an agreement error. The parser spreads activation to every word that has features +SUBJECT and +PLURAL. Now, both the subject "key" and the attractor "cells" fail to fully match the retrieval cues, each has only one matching feature, +SUBJECT or +PLURAL. The attractor and the subject therefore receive the same amount of activation, and noise in the system decides which word is retrieved.

An important difference to the production accounts discussed above is that the Lewis and Vasishth model predicts attraction effects in ungrammatical sentences irrespective of their syntactic structure—any noun, regardless of position and syntactic role can be retrieved instead of the subject as long as it matches sufficiently many retrieval cues. The Lewis and Vasishth model can also explain the increase in attraction error rates when the attractor superficially resembles the sentential subject (Engelmann et al., 2019), for instance, when the attractor's case marking is ambiguous between nominative and the actual case (Avetisyan et al., 2020; Badecker & Kuminiak, 2007; Hartsuiker, Schriefers, Bock, & Kikstra, 2003; Slioussar & Malko, 2016).

Unlike the production accounts, the Lewis and Vasishth model cannot explain the singular–plural asymmetry present in many studies, that is, more attraction errors when the subject noun is singular and the attractor noun is plural than in the reverse configuration. The reason is that, unlike Feature Percolation, the Lewis and Vasishth model assumes that singular is marked just as plural.[2] Similar asymmetries, which the Lewis and Vasishth model also cannot explain, have been found in gender (more errors in sentences with a masculine subject noun and feminine attractor that the other way around in Slovak and Russian, see Badecker & Kuminiak, 2007; Slioussar & Malko, 2016).[3]

The Lewis and Vasishth model has also been invoked to account for attraction effects in production (Badecker & Kuminiak, 2007; Konieczny, Schimke, & Hemforth, 2004). The idea is that, to build syntactic structure for production, we need to keep in memory what has already been said and what we are planning to say, and that the memory substrate used in this process is likely the same as for comprehension.

### *1.3. Differences between production and comprehension accounts*

Production and comprehension accounts differ not only in the mode of language use, they also postulate fundamentally different mechanisms. According to the Lewis and Vasishth model, attraction arises at the moment of dependency formation—an incorrect syntactic chunk is retrieved to form the dependency—whereas according to both Feature Percolation and Marking and Morphing, attraction arises due to the incorrect encoding of the subject's number. As a consequence, the Lewis and Vasishth model predicts that, if attraction occurs, the attractor noun is perceived as the subject, while the production accounts both predict that

the subject noun is identified correctly, just with incorrect number marking. The available evidence is inconclusive so far but favors the production accounts because it suggests that the attractor noun is perceived as the subject only in a minority of cases (Patson & Husband, 2016, Schlueter et al., 2019).

The two classes of accounts further differ in the role they assign to semantic information. Marking and Morphing allows a narrow set of semantic properties, either of the subject noun, such as conceptual plurality, or of the whole noun phrase, such as distributivity or conceptual number (Schlueter, Williams, & Lau, 2018), to influence whether agreement attraction occurs. Other semantic features are not assumed to be involved in attraction. This means that Marking and Morphing cannot account for the increase in attraction rates due to the goodness of thematic fit between the attractor and the verb (Thornton & MacDonald, 2003), or due to higher semantic integration between the subject and the attractor within the noun phrase, such that the subject and the attractor are conceptualized as a whole ("the painting with the flowers" as opposed to "the painting of the flowers," where flowers exist independently, see Solomon & Pearlmutter, 2004), or due to attractor being an animate noun (Bock & Miller, 1991, Experiment 3). Many of these semantic influences on agreement computation are easily explained by the Lewis and Vasishth model since in the Lewis and Vasishth model, semantic features receive the same treatment and have the same weight as all other types of features, including morphosyntactic. Crucially, the role of semantic features is therefore not limited to influencing agreement computations and therefore modulating agreement attraction. As a consequence, the Lewis and Vasishth model makes a surprising prediction, namely, that attraction effects should occur not only in (morpho)syntax and agreement, but also in other linguistic domains. As Lewis and Vasishth state (2005, p. 411):

> In this model, we have realized only syntactic cues, which are used primarily to reactivate predicted structure to unify with. However, the model can accommodate a richer set of cues – for example, there may also be semantic cues derived from specific lexical constraints (e.g., the semantic constraints that a verb places on its subject).

This means that the Lewis and Vasishth model predicts not only agreement attraction errors, but also analogous *semantic attraction errors*. These could be reflected in the acceptance of a verb that thematically fits the attractor, but not the subject noun. For example, in the sentence "The drawer with the knife cuts …," the semantic attractor "knife" satisfies the semantic restrictions set by the verb "cuts" better than the subject noun "drawer." In comprehension, the verb "cuts" should therefore be easier to process in the presence of an attractor that can perform the cutting action than in the presence of an attractor that cannot, such as "handle" in "The drawer with the handle cuts …." These effects would precisely mirror agreement attraction effects, but crucially, they would arise independently of morphosyntactic computations, that is, even in sentences that are morphosyntactically well-formed.

Note that this proposal differs from the one made in Thornton and MacDonald (2003) where semantic features were shown to influence agreement computations. While relevant in the present context, the proposal that plausibility directly affects morphosyntactic agreement is more narrow in scope than the idea of purely semantic attraction effects predicted by the

Lewis and Vasishth model. Thus, finding evidence for purely semantic attraction effects independent of morphosyntactic computations would show that attraction is potentially a much broader phenomenon than previously believed and that semantic features not only modulate morphosyntactic computations but also give rise to their own non-morphosyntactic attraction effects.

Semantic attraction as sketched above is conceptually related to semantic interference, which has been repeatedly demonstrated in grammatical sentences (Van Dyke, 2007; Van Dyke, Johns, & Kukona, 2014; Van Dyke & McElree, 2006, 2011). However, the processing of well-formed sentences likely differs in important ways from the processing of ill-formed sentences investigated here. To our knowledge, there is only one study that provides initial evidence for semantic interference in ill-formed sentences, that is, for what we refer to as semantic attraction. In two eye-tracking experiments, Cunnings and Sturt (2018) presented participants with sentences like (6):

(6) a. Sue remembered the letter that the butler with the *cup* accidentally shattered.
b. Sue remembered the letter that the butler with the *tie* accidentally shattered.

Both sentences are implausible, but Cunnings and Sturt found that the verb "shattered" was processed faster in condition (2-a), where the local non-subject noun "cup" was semantically a good fit for the verb "shattered," than in (2-b), where the local noun was "tie," that is, an object that cannot be shattered. These results resemble agreement attraction effects, but it is not clear that these semantic effects are necessarily subserved by the same mechanisms. If semantic attraction effects exist at all, they should, according to the Lewis and Vasishth model, have the same size as agreement attraction effects (since the model treats all types of features the same way).

The purpose of the present study therefore is first to conceptually replicate the semantic attraction effect demonstrated by Cunnings and Sturt (2018) using a different experimental paradigm, and second, to build on their work by examining more closely whether the effects they observed constitute genuine attraction effects, that is, effects plausibly arising from the same mechanisms underlying agreement attraction. To do so, we compared configurations with semantic attraction and morphosyntactic attraction side by side using a modified version of the forced-choice paradigm that has been extensively used to study agreement attraction.

The goal of Experiment 1 was to establish whether semantic attraction errors occur in this paradigm and, if yes, whether the rate of semantic attraction errors is similar to that of morphosyntactic (agreement) attraction errors as predicted by Lewis and Vasishth model. Experiment 2 replicated the findings of Experiment 1 and included two additional conditions that give us further insight into whether and how semantic and morphosyntactic attractions interact when they occur simultaneously. To address a potential confound in Experiments 1 and 2, Experiment 3 replicated the results of Experiments 1 and 2 with an improved set of stimuli and a slightly adapted procedure. Finally, we report computational simulations with the Lewis and Vasishth model to see whether the model captures the finer-grained quantitative patterns in the data.

### 1.4. Disclosures

All reported studies have been carried out in accordance with the Declaration of Helsinki. All participants provided informed consent. The full list of materials used in all reported experiments, the collected data, and analysis code are available from the project page at the Open Science Framework.[4] The full list of materials is also provided in the Supporting Information.

## 2. Experiment 1

The goal of Experiment 1 was to demonstrate semantic attraction and to compare it to classic agreement attraction. The method we used was a modified version of a forced-choice completion task. The classic version of the task presents participants with a sentence preamble and prompts them to choose one out of two verbs as a plausible continuation. Instead of two verbs, we showed only one and asked participants to judge whether or not it was a plausible continuation of the preamble. If the rate of mistakes is increased when the attractor matches the verb thematically (with the verb being a good morphosyntactic fit for the subject), that would constitute a purely semantic attraction effect.

A secondary goal was to compare semantic attraction effects to morphosyntactic (i.e., agreement) attraction effects: Are they equally sized? And how do semantic and morphosyntactic attraction interact? Are the effects of morphosyntactic and semantic attraction additive, or under-, or superadditive? If there is evidence for an interaction, this might favor a common underlying substrate (Roberts & Sternberg, 1993; Sternberg, 1998), while a lack of interaction would be consistent both with a single common and independent underlying substrates. Furthermore, establishing whether or not the effects are additive will constrain computational cognitive models of attraction.

### 2.1. Methods

#### 2.1.1. Participants

Prior to running the experiments, we conducted a prospective power analysis using simulated data. The estimated power to detect a 5% attraction effect with 200 observations per condition was between 60% and 80%. Due to limited funds, we collected only approximately 179 observations per condition.

Participants ($N = 1{,}100$) were recruited on Prolific, a crowd-sourcing platform for academic studies. Participants were prescreened to be self-reported native speakers of English who were born in the United States or the United Kingdom, citizens of the United States or the United Kingdom, and residents of the United States or the United Kingdom at the time of participation. We chose to conduct the experiment in English because this gave us access to a wider population of participants necessary for a single-trial experiment. The second reason was that the control effect, morphosyntactic attraction, is well established in English, and might be smaller in other languages, such as those with case marking (Avetisyan et al., 2020). Participation took approximately one minute and was compensated

Table 1
Example experimental item

| Condition | | | Violation | Attraction |
|---|---|---|---|---|
| a. | The drawer with the handle | OPEN | Morphosyntactic | None |
| b. | The drawer with the handles | OPEN | Morphosyntactic | Morphosyntactic |
| c. | The drawer with the handle | CUTS | Semantic | None |
| d. | The drawer with the knife | CUTS | Semantic | Semantic |
| e. | The drawer with the handle | CUT | Double | None |
| f. | The drawer with the knives | CUT | Double | Double |
| g. | The drawer with the knife | CUT | Double | Semantic |
| h. | The drawer with the handles | CUT | Double | Morphosyntactic |

Conditions (a–f) were tested in Experiment 1, and conditions (g) and (h) were added in Experiment 2. "Double" stands for simultaneous morphosyntactic and semantic attraction or violation.

with 10p (0.1 GBP). After finishing the experimental task, participants had to indicate (again) whether they were native speakers of English, U.S./UK citizens, and that they spent the first 5 years of their life in the United States or the United Kingdom. After excluding data from participants who responded negatively to any of these questions, 1,072 individuals were left for the analysis.

### *2.1.2. Materials*

We tested 25 item sets (see Table 1, conditions a–f) in which the verb never fully matched the subject. It mismatched either the subject's number (morphosyntactic violation), or theme (semantic violation), or both (double violation). At the same time, the verb could match or mismatch the attractor in number (morphosyntactic attraction), meaning (semantic attraction), or both (double attraction). This set of conditions allowed us to test morphosyntactic attraction (conditions b vs. a), semantic attraction (d vs. c), as well as double attraction (f vs. e).

The items had the following structure: The subject noun was followed by a prepositional phrase containing the attractor. The verb had clear thematic restrictions that allowed for only a subset of nouns to plausibly serve as a subject. Subject- and attractor-verb combinations were created with the aim to avoid metonymic and metaphorical sense transfers (e.g., a person glowing with joy).

To assess the degree of semantic match/mismatch, we conducted a post-hoc semantic similarity assessment. For each of the 25 item sets, we constructed five subject–verb pairs using the three noun phrases (the subject and the two attractors) and the two verbs as follows: "The drawer/the handle opens," "The drawer/the handle/the knife cuts." For each pair, we obtained latent semantic analysis (LSA) measures using the lsa.colorado.edu website (the pair-wise comparison tab).[5] We analyzed the LSA measures using Bayesian linear regression (for details, see Appendix A). As expected, the preambles constructed to be plausible had systematically higher LSA values than the ones constructed to be implausible.

### *2.1.3. Procedure*

The study was conducted as a single-trial online experiment in which each participant saw only one item and only one of the experimental conditions. This avoided adaptation of

processing strategies to the stimuli and task. We discuss the implications of the single-trial design, specifically with respect to statistical power, in the general discussion.

The experiment consisted of instructions, the experimental probe (see Table 1), and the debriefing questions mentioned above (native language, citizenship, country of residence during the first 5 years of life). The instruction was: "In this experiment, we ask you to decide whether a word fits into the sentence or not. First, you will see the word in question. Memorize it and proceed when you are ready. Next, you will see a sentence fragment. Decide whether the word that you memorized could be the next word in this sentence. Click on the red cross if you think that the word does not fit (for whatever reason) and the green check mark if you think that the word does fit." Thornton and MacDonald (2003) showed that presenting the verb before the preamble produces the same results as the more common oral production task where the verb follows the preamble.

To indicate whether the verb was a possible continuation of the sentence, participants had either to click on one of the symbols (green check mark or red X mark) or to press 1 or 2 on the keyboard, where 1 corresponded to "good fit" and 2 to "bad fit." Note that the verb never perfectly matched the subject, and the correct response was therefore always to reject the verb. However, since each participant performed only one trial, the correct response could not be guessed based on knowledge from previous trials.

A reviewer pointed out that participants might judge the fit of the subject and the verb even before seeing the attractor immediately upon seeing the subject, that is, before even reading the attractor. This strategy is, however, unlikely: Participants did not know that the word in question was a verb ("cut" could be a verb or a noun). And even if they assumed that it was a verb, it was not clear that the first noun was necessarily the corresponding subject. Participants therefore needed to read the full preamble to do the task, and, to preview, the results indicated that they did. Having said that, in a standard repeated measures design, participants may well have developed such a strategy, which highlights one of the benefits of the single-trial design used here.

The experiment was programmed using the Ibex[6] software and run on the IbexFarm cloud service operated by Alex Drummond.

### *2.1.4. Data analysis*

All analyses were conducted with the R system for statistical computing (R Development Core Team, 2009). Accuracy in the judgment tasks was analyzed using hierarchical logistic regressions in the Bayesian framework (Vasishth, Nicenboim, Beckman, Li, & Kong, 2018) using the "brms" package (Bürkner, 2017). Plots were produced with the "ggplot2" and "tidybayes" packages (Kay, 2019; Wickham, 2016). Inferences were based on the posterior distributions of the parameters, which are reported in terms of the posterior mean, the 95% percentile interval (CrI), and the posterior probability that the true parameter value lies on one side of zero (e.g., $P(\beta > 0) = 0.9$).[7] When nearly all of the posterior mass for an estimate fell on one side of zero, we considered the effect reliable. However, note that we do not adopt a strict threshold here, we instead evaluate the posterior in a graded fashion.

Treatment contrasts were used to code the two factors: the type of violation (morphosyntactic, semantic, or both) with morphosyntactic violations serving as the reference level,

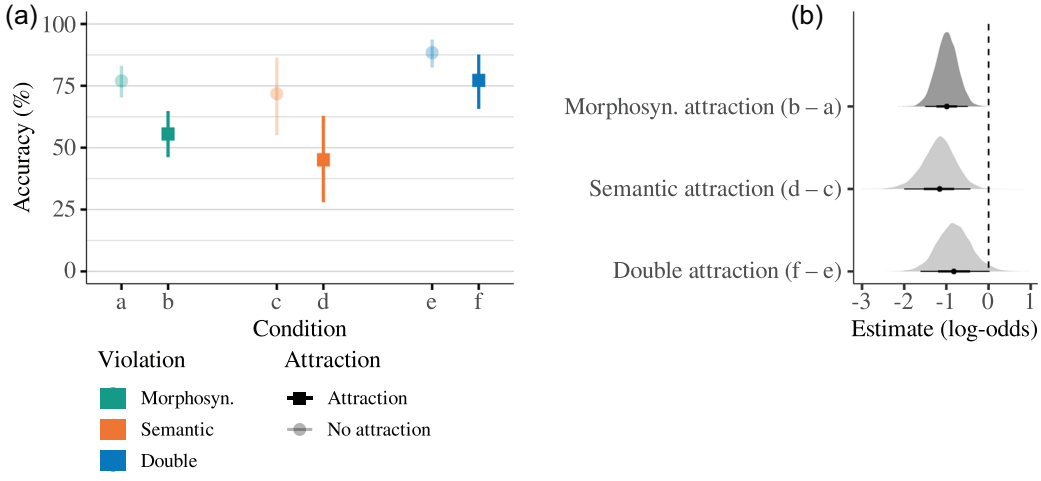


Fig. 1. Results of experiment 1. Panel A: Estimated condition means with 95% credible intervals. Panel B: Posterior distributions for the attraction effects. The posterior for the semantic and the double attraction effects (light gray) was obtained by combining the posteriors for morphosyntactic attraction and the posterior for the difference between the morphosyntactic and semantic attraction or the morphosyntactic and the double attraction. All parameters are on the log-odds scale. Error bars around the posterior means represent 66% (thick) and 95% (thin) credible intervals.

and attraction (none, morphosyntactic, semantic) with no attraction being the reference level (Schad, Vasishth, Hohenstein, & Kliegl, 2020). We estimated both simple effects as well as the interactions.[8] As regularizing priors for the fixed effects, we used a normal distribution with mean 0 and standard deviation 1. These priors allow for a wide range for effect sizes from 0% up to 90% and discourage only implausibly large effects bigger than 90%. For other parameters, we used default brms priors. Random effects for participants were not needed since each participant contributed only one measurement (single-trial design). Random effects for items were maximal with full variance–covariance matrices (Barr, Levy, Scheepers, & Tily, 2013; Schielzeth & Forstmeier, 2009). The models were run with four chains, 4,000 iterations per chain, with half of iterations discarded as warm-up. All $\hat{R}$s were close to 1, indicating that the chains mixed properly.

### 2.2. Results

The estimated proportions of correct responses in each condition are shown in Fig. 1a, and the posterior distributions of the parameters in Fig. 1b.

Accuracy in condition (a), the baseline for morphosyntactic attraction, was 77% ($\hat{\beta} = 1.22$, 95%-CrI: [0.86, 1.60]). There was little evidence suggesting that accuracy in the baseline for semantic attraction (c) differed from the baseline for morphosyntactic attraction (a) (77% vs.73%, $\hat{\beta} = 0.25$, 95%-CrI: [−1.13, 0.64], $P(\beta < 0) = 0.73$). However, accuracy in condition (e), the baseline for double attraction, was reliably higher than that in (a) (77% vs. 89%, $\hat{\beta} = 0.85$, 95%-CrI: [0.21, 1.57], $P(\beta < 0) = 0.004$), which suggests that double

subject–verb fit violations were easier to spot than isolated morphosyntactic or semantic violations.

We found the classic agreement attraction effect, that is, accuracy was considerably lower in condition (b) with morphosyntactic attraction compared to baseline (a) without attraction (77% vs. 56%, $\hat{\beta} = -1.00$, 95%-CrI: $[-1.50, -0.49]$, $P(\beta < 0) = 0.999$). Neither semantic nor double attraction effects differed from the morphosyntactic attraction effect (semantic attraction: 49% vs. 45%, $\hat{\beta} = -0.17$, 95%-CrI: $[-1.07, 0.68]$, $P(\beta < 0) = 0.65$; double attraction: 75% vs. 78%, $\hat{\beta} = 0.17$, 95%-CrI: $[-0.75, 1.12]$, $P(\beta < 0) = 0.35$).[9]

To assess more directly whether semantic attraction was reliably different from zero, we combined the posterior of the morphosyntactic attraction effect with the posterior of the difference between the morphosyntactic and semantic attraction effects (McElreath, 2016). The resulting posterior for the size of the semantic attraction effect (comparison between conditions c and d) suggested a highly reliable decrease in response accuracy in the presence of semantic attraction (73% vs. 45%, $\hat{\beta} = -1.17$, 95%-CrI: $[-1.96, -0.47]$, $P(\beta < 0) = 0.999$).

These effect sizes are slightly bigger (potentially due to the single-trial design) but largely in line with those reported in earlier research using similar tasks: 17% in Schlueter et al. (2019), 18% in Staub (2009), and 13% and 19% in the sentence repetition paradigm used by Thornton and MacDonald (2003).

### 2.3. Discussion

In line with the predictions of the Lewis and Vasishth model, we found a semantic attraction effect similar in manifestation and size to the classic morphosyntactic attraction effect. These results conceptually replicate the semantic interference effect reported by Cunnings and Sturt (2018) in eye movements during reading. We will review the broader implications of this finding in the general discussion.

Another goal of Experiment 1 was to assess whether these two types of attraction effects interact: under- or overadditive effects would favor a common underlying mechanism. While we observed an interaction—the effect of double attraction was not larger than single morphosyntactic or semantic attraction (in log-odds)—the relevant comparison of conditions may have been flawed: Isolated morphosyntactic and semantic attraction effects were tested with subject–verb combinations that violated either morphosyntactic agreement or semantic plausibility. In contrast, double attraction was tested with subject–verb combinations that mismatched along both dimensions, morphosyntax *and* semantic plausibility. The results show that this double violation was easier to spot than single violations (higher accuracy in condition e than in conditions a or c). So, while double attraction might be stronger than single attraction, that effect may have been partly counteracted and canceled out by the easier detection of the subject–verb mismatch in (e).

To address this shortcoming of the design, we conducted Experiment 2 with two additional conditions. In these conditions, both morphosyntactic and thematic fit between the subject noun and the verb were violated (as in conditions e and f). In each of the additional conditions, the attractor matched the verb along a single dimension, either morphosyntactic

or semantic. Experiment 2 therefore allows us to cleanly compare morphosyntactic, semantic, and double attraction in the presence of the same double subject–verb fit violation. A secondary goal of Experiment 2 was to replicate the semantic attraction effect found in Experiment 1.

## 3. Experiment 2

### *3.1. Methods*

#### *3.1.1. Participants*

Participant recruitment procedure and exclusion criteria were the same as for Experiment 1. Individuals who participated in Experiment 1 were blocked from participating in Experiment 2. We tested more participants in order to maintain the same number of observations per condition as in Experiment 1 and thus the same statistical power: 1,450 individuals took part in the experiment; after applying exclusion criteria, data from 1,426 individuals were left in the analysis.

#### *3.1.2. Materials*

We retained all conditions from Experiment 1 but added conditions (g) and (h) that introduce morphosyntactic and semantic attraction manipulation in the presence of a double violation of subject–verb fit (see Table 1).

#### *3.1.3. Procedure*

Experimental procedure was identical to that of Experiment 1.

#### *3.1.4. Data analysis*

To establish the reliability of the semantic attraction effect, we repeated the analysis from Experiment 1 but excluded conditions (e) and (f), since comparisons with these conditions are confounded as explained above. Consequently, for the analysis of conditions (a)–(d), we were left with a $2 \times 2$ design with factors *type of violation* (morphosyntactic or semantic) and *attraction* (present or not). As in Experiment 1, these factors were coded as treatment contrasts with morphosyntactic violation as the reference level for factor *type of violation* and no attraction as the reference level for the factor *attraction*.

To assess the interaction of morphosyntactic and semantic attraction in conditions (e)–(h), we fit a separate model with factors *morphosyntactic attraction*, *semantic attraction*, and their interaction. Morphosyntactic and semantic attraction were coded with sum contrasts such that the parameter estimates captured the main effects of morphosyntactic and semantic attraction (i.e., the effect averaged across the levels of the respective other factor). As before, the models included full by-item random effects and used the same regularizing priors.

### *3.2. Results*

The estimated proportions of correct responses in each condition can be seen in Fig. 2a, and posterior distributions of the parameters in Figs. 2b and 2c.

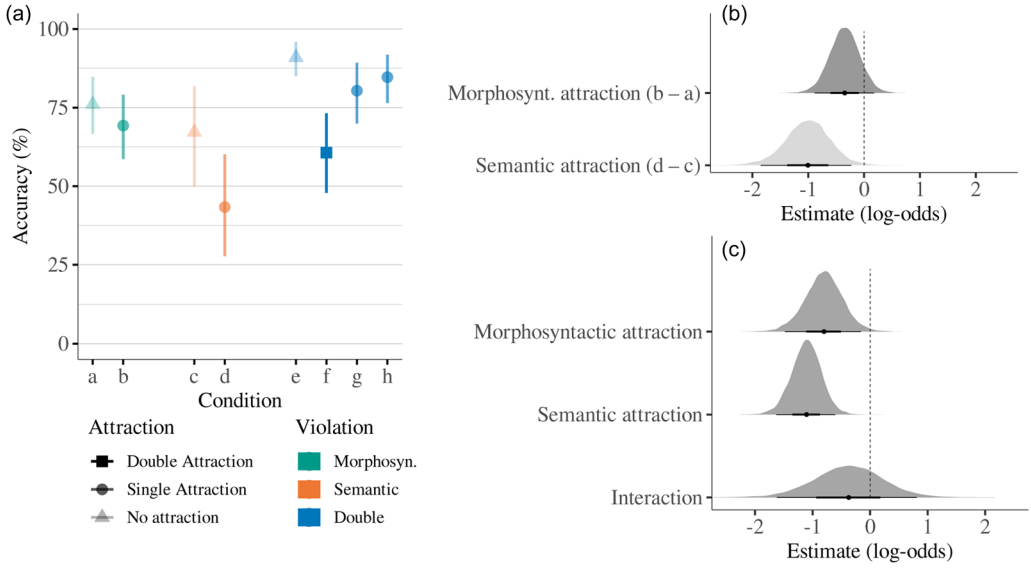


Fig. 2. Results of experiment 2. Panel A: Estimated condition means with 95% credible intervals. Panel B: Posterior distributions for the model of conditions (a)–(d). The posterior for semantic attraction (light gray) was obtained by combining the posteriors for morphosyntactic attraction and the difference between the semantic and morphosyntactic attraction. Panel C: Posterior distributions for the model of conditions (e)–(h). All parameters are on the log-odds scale. Error bars around the posterior means represent 66% (thick) and 95% (thin) credible intervals.

*Analysis replicating results of Experiment 1 (conditions a–d)* Accuracy in the baseline condition for morphosyntactic attraction (a) was 76% ($\hat{\beta} = 1.18$, 95%-CrI: [0.69, 1.72]). Accuracy in the baseline condition for semantic attraction (c) was slightly lower but not reliably so (76% vs. 67%, $\hat{\beta} = -0.45$, 95%-CrI: [−1.37, 0.48], $P(\beta < 0) = 0.83$). The morphosyntactic attraction effect (a vs. b) was in the expected direction but not entirely reliable this time (76% vs. 70%, $\hat{\beta} = -0.36$, 95%-CrI: [−0.88, 0.17], $P(\beta < 0) = 0.90$). The effect of semantic attraction was numerically bigger but did not reliably differ from the effect of morphosyntactic attraction (59% vs. 43%, $\hat{\beta} = -0.65$, 95%-CrI: [−1.63, 0.27], $P(\beta > 0) = 0.93$). As in Experiment 1, we combined posteriors to get a direct estimate of the semantic attraction effect (d vs. c) and to see whether it was different from zero. The result shows that semantic attraction reliably decreased response accuracy (67% vs. 43%, $\hat{\beta} = -1.01$, 95%-CrI: [−1.83, −0.24], $P(\beta < 0) = 0.993$), thus replicating the semantic attraction effect found in Experiment 1.

*Analysis testing the interaction of morphosyntactic and semantic attraction (conditions e–h)* The average accuracy across conditions was 82% ($\hat{\beta} = 1.5$, 95%-CrI: [1.1, 1.9]). Morphosyntactic attraction reliably decreased response accuracy (87% vs. 75%, $\hat{\beta} = -0.8$, 95%-CrI: [−1.5, −0.17], $P(\beta < 0) = 0.99$). Likewise, semantic attraction reliably decreased accuracy (89% vs. 72%, $\hat{\beta} = -1.1$, 95%-CrI: [−1.6, −0.63], $P(\beta < 0) = 0.999$). There was no

interaction of morphosyntactic and semantic attraction, that is, their effects were approximately additive in log-odds (83% vs. 80%, $\hat{\beta} = -0.37$, 95%-CrI: $[-1.6, 0.82]$, $P(\beta < 0) = 0.74$).

### 3.3. Discussion

The two goals of Experiment 2 were to confirm the reliability of the semantic attraction effect and to test whether the semantic and morphosyntactic attraction effects are additive given appropriate control conditions. We successfully replicated the semantic attraction effect, both in the context of single and double subject–verb fit violations. The outcomes of Experiment 2 also suggest that morphosyntactic and semantic attraction effects are approximately additive, which is consistent with a single common substrate but also with separate substrates for morphosyntactic and semantic attraction.

While these results are in line with the key predictions of the Lewis and Vasishth model, the design of our experimental items has a potential confound that could, in principle, account for the semantic attraction effect: In semantic attraction conditions with single subject–verb fit violations (d), the attractor and the verb could sometimes form locally coherent noun–noun compounds, such as "tree blossoms," "knife cuts," "fountain bubbles," and so on. Thus, it is possible that participants accepted the memorized word as a continuation in condition (d) not due to semantic attraction but because they adopted a noun–noun compound interpretation. This is possible in particular because we did not instruct participants to interpret the continuation word as a verb. To assess whether and how much this confound may have influenced our estimate of the semantic attraction effect, we reran the analysis with the data from both experiments but excluded the stimuli that allowed the compound interpretation ($N = 1{,}426$, exactly as many as the original Experiment 2). In this analysis, the key semantic attraction effect was still present in conditions with both single and double subject–verb fit violations (see Appendix C for details). To get a better and unbiased estimate of semantic attraction, we also replicated Experiment 2 using a new set of items in which the noun–noun compound interpretation was ruled out.

## 4. Experiment 3

### 4.1. Methods

#### 4.1.1. Participants

Participant recruitment procedure was the same as for Experiments 1 and 2. Individuals who participated in Experiments 1 and 2, as well as individuals who participated in the pretest of experimental materials, were blocked from participating in Experiment 3. We tested 2,600 participants; after applying exclusion criteria, data from 2,454 participants were left in the analysis (compare to the pooled $N = 2{,}498$ in Experiments 1 and 2).

#### 4.1.2. Materials

We created a new set of experimental items. To exclude the possibility of noun–noun compound interpretations, the attractor noun was followed by an adverb unambiguously signaling that the memorized word must be a verb, see example item in Table 2.

Table 2
Example experimental item from Experiment 3

| Condition | | | Violation | Attraction |
|---|---|---|---|---|
| a. | The bakery near the office building rarely | SMELL | Morph. | None |
| b. | The bakery near the office buildings rarely | SMELL | Morph. | Morph. |
| c. | The bakery near the office building rarely | SPRAYS | Semantic | None |
| d. | The bakery near the fire hydrant rarely | SPRAYS | Semantic | Semantic |
| e. | The bakery near the office building rarely | SPRAY | Double | None |
| f. | The bakery near the fire hydrants rarely | SPRAY | Double | Double |
| g. | The bakery near the fire hydrant rarely | SPRAY | Double | Semantic |
| h. | The bakery near the office buildings rarely | SPRAY | Double | Morph. |

"Double" stands for simultaneous morphosyntactic and semantic attraction or violation, depending on the column.

To ensure that the semantic match/mismatch was actually perceived as such by native English speakers, we conducted a plausibility norming pretest. For each of the 32 items sets, we constructed five sentence preambles using the three noun phrases (the subject and the two attractors) and the two verbs as follows: "The bakery/the office building rarely smells …", "The bakery/the office building/the fire hydrant rarely sprays …". Participants then rated these preambles on a 1–7 Likert scale. Each participant ($N = 50$, recruited online on Prolific) saw all 32 experimental items, each item in one out of five conditions. Lists were created following a Latin square design. We analyzed the results using Bayesian ordinal regression (Veríssimo, 2021; for details, see Appendix B). As expected, the preambles constructed to be plausible received systematically higher ratings than the ones constructed to be implausible. Based on model estimates, we excluded four items for which the estimated difference between the plausible and implausible conditions was the smallest and not reliably different from zero. We additionally excluded one item for which a distributive interpretation of the number attraction condition was available. This left us with 27 experimental items for Experiment 3. The full list of experimental items can be found in the Supporting Information.

#### *4.1.3. Procedure*

The experimental procedure was similar to that of Experiments 1 and 2 with a small modification: We introduced two training sentences so that participants could familiarize themselves with the experimental procedure. For the training sentences, participants also had to memorize a word (which was not necessarily a verb) and judge whether the word fit the sentence preamble. One of the training sentences was ill-formed ("The house by the new FURIOUSLY …)", and we excluded data from participants who failed to notice the ill-formedness. This lead to exclusion of 5% of data points, but the results remained the same when data from these participants were retained.

#### *4.1.4. Data analysis*

We replicated both analyses from Experiment 2.

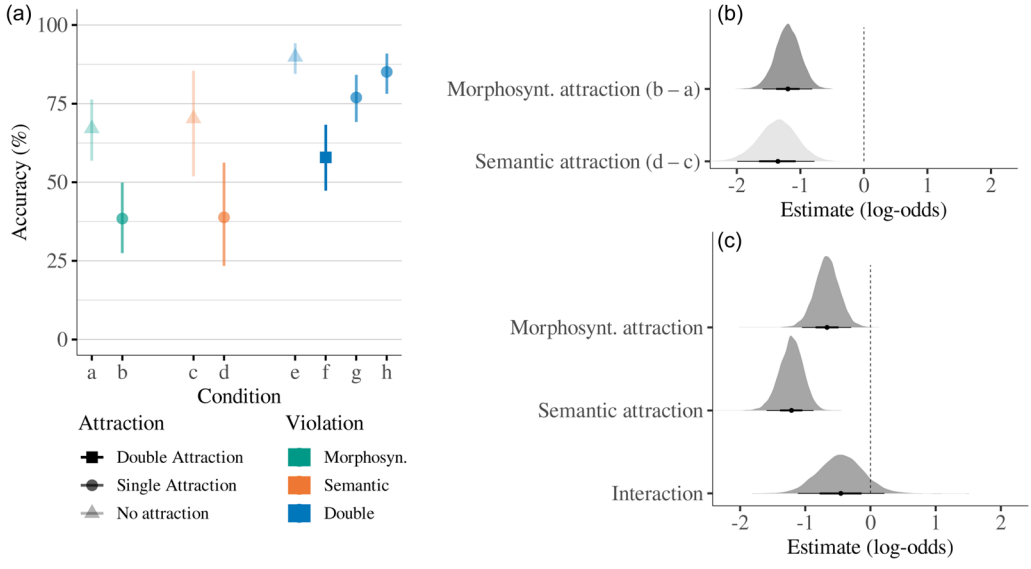


Fig. 3. Results of experiment 3. Panel A: Estimated condition means with 95% credible intervals. Panel B: Posterior distributions for the model of conditions (a)–(d). The posterior for semantic attraction (light gray) was obtained by combining the posteriors for morphosyntactic attraction and the difference between the semantic and morphosyntactic attraction. Panel C: Posterior distributions for the model of conditions (e)–(h). All parameters are on the log-odds scale. Error bars around the posterior means represent 66% (thick) and 95% (thin) credible intervals.

### *4.2. Results*

The estimated proportions of correct responses in each condition can be seen in Fig. 3a, and posterior distributions of the parameters in Figs. 3b and 3c.

*Conditions a–d* Accuracy in the baseline condition for morphosyntactic attraction (a) was 67% ($\hat{\beta} = 0.72$, 95%-CrI: [0.28, 1.2]). Accuracy in the baseline condition for semantic attraction (c) was only slightly higher (67% vs. 71%, $\hat{\beta} = 0.18$, 95%-CrI: [−0.81, 1.2], $P(\beta < 0) = 0.37$). The expected morphosyntactic attraction effect was highly reliable (a vs. b) (67% vs. 38%, $\hat{\beta} = -1.2$, 95%-CrI: [−1.6, −0.81], $P(\beta < 0) = 0.999$). The effect of semantic attraction was numerically bigger but did not reliably differ from the effect of morphosyntactic attraction (42% vs. 38%, $\hat{\beta} = -0.17$, 95%-CrI: [−0.9, 0.55], $P(\beta < 0) = 0.67$). As in Experiment 1, we combined posteriors to get a direct estimate of the semantic attraction effect (d vs. c). The result shows that semantic attraction reliably decreased response accuracy (39% vs. 70%, $\hat{\beta} = -1.36$, 95%-CrI: [−1.99, −0.78], $P(\beta < 0) = 0.999$) thus replicating the semantic attraction effect found in Experiments 1 and 2. Note that both attraction effects, morphosyntactic and semantic, were larger in Experiment 3 than in Experiments 1 and 2.

*Conditions e–h* The average accuracy across conditions was 80%, ($\hat{\beta} = 1.4$, 95%-CrI: [1, 1.8]). Morphosyntactic attraction reliably decreased response accuracy (85% vs. 74%,

$\hat{\beta} = -0.67$, 95%-CrI: $[-1.1, -0.3]$, $P(\beta < 0) = 0.999$). Likewise semantic attraction reliably decreased accuracy (88% vs. 68%, $\hat{\beta} = -1.2$, 95%-CrI: $[-1.6, -0.87]$, $P(\beta < 0) = 0.999$). There was a small numerical trend toward a superadditive interaction of morphosyntactic and semantic attraction (82% vs. 78%, $\hat{\beta} = -0.46$, 95%-CrI: $[-1.1, 0.21]$, $P(\beta < 0) = 0.91$). Again, attraction effects were numerically larger in Experiment 3 than in Experiment 2.

To obtain an even more precise estimate of the interaction, we also combined data from all three experiments and repeated the last analysis (the possible confound in the stimulus design in Experiments 1 and 2 did not affect the relevant conditions). The analysis of the combined dataset ($N = 2{,}338$) still did not yield reliable evidence for an interaction between morphosyntactic and semantic attraction (82% vs. 80%, $\hat{\beta} = -0.27$, 95%-CrI: $[-0.78, 0.24]$, $P(\beta < 0) = 0.85$).

### *4.3. Discussion*

The main goal of Experiment 3 was to replicate the results of Experiment 2 with an improved set of stimuli that rule out noun–noun compound interpretations in one of the two semantic attraction conditions (d). All effects reported in previous experiments were qualitatively replicated. We found semantic attraction effects both in single and double subject–verb fit violation configurations, and the effect size of semantic attraction was similar to that of morphosyntactic attraction. These results are qualitatively consistent with the predictions of Lewis and Vasishth model as outlined in the introduction. In the following section, we investigate whether the Lewis and Vasishth model also provides a good quantitative fit to the data.

## 5. Computational simulation with the Lewis and Vasishth model

In the following, we explain the predictions of the Lewis and Vasishth model in more detail and investigate whether these predictions can be improved by adjusting the model parameters. For this purpose, we used an implementation of the Lewis and Vasishth model in R, the so-called interACT model (Engelmann et al., 2019).

We first lay out the linking hypothesis that allows us to link model dynamics to the response variable produced by our task. With default parameters, the Lewis and Vasishth model predicts the resulting parse and the time it will take to build this parse. The model does not explicitly track sentence grammaticality or well-formedness: a syntactic structure is either built, in which case it is assumed to be correct, or it is not built if retrieval of a constituent from memory fails. The most straightforward mapping from model dynamics to our task is therefore to assume that a failure to build a structure results in rejecting the sentence as ill-formed (correct response in our task), and that retrieving a noun from memory and subsequent formation of a subject–verb dependency—correct or not—results in accepting the verb (incorrect response in our task). The first scenario corresponds to failing to build a parse—after all, there is no correct parse—whereas the second scenario corresponds to the illusion of a correct parse when there is none.

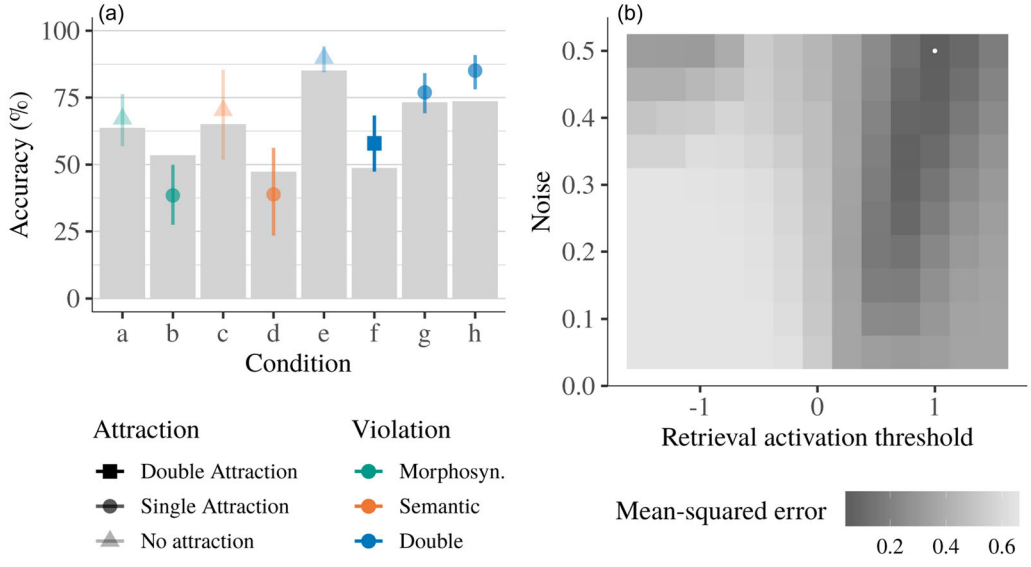


Fig. 4. Modeling results. Panel A: Model predictions compared to the observed data. Gray bars represent predictions of Lewis and Vasishth model with the best-fitting set of parameters. Colored lines represent the 95% credible intervals of the observed condition means from Experiment 3. Panel B: Prediction error as a function of the two varied parameters. Dark shades indicate a better fit. The white dot marks the best-fitting parameter combination.

Recall that retrieval failure happens when the activations of all chunks in memory lie below the retrieval threshold—the lower the activation of each chunk, the higher the probability of retrieval failure, and therefore, of a correct response. In all attraction conditions, the attractor matches more retrieval cues than in the respective control conditions, which increases the activation of the attractor noun and the probability that it will be retrieved and attached. Therefore, in attraction conditions, the probability of a correct response is always predicted to be lower.

Furthermore, retrieval failure (leading to correct responses) should happen more often in conditions with double violation of subject–verb fit than in conditions with single violation of subject–verb fit. The Lewis and Vasishth model therefore predicts higher accuracy in conditions with double violation of subject–verb fit (the exception is condition f where the attractor matches two features of the verb).

While it appears that the Lewis and Vasishth model should in principle reproduce the qualitative pattern of our results (see Fig. 4a), the standard version of the model, as reported by Lewis and Vasishth (2005), has a retrieval activation threshold that is set so low that some item will always be retrieved from memory even if it mismatches two out of three retrieval cues. Consequently, the Lewis and Vasishth model with default parameter settings predicts no failed retrievals, and hence 0% accuracy in all conditions, which is clearly implausible. To address the default model's failure to capture the observed pattern, we explored through

Table 3
Summary of cue–feature matches for the subject and the attractor nouns across experimental conditions predicted by Lewis and Vasishth model

| Condition | | | Retrieval cues | Subject matches | Attractor matches |
|---|---|---|---|---|---|
| a. | The drawer with the handle | OPEN | +SUBJ<br>+PL<br>+OPENABLE | +SUBJ<br><br>+OPENABLE | |
| b. | The drawer with the handles | OPEN | +SUBJ<br>+PL<br>+OPENABLE | +SUBJ<br><br>+OPENABLE | +PL |
| c. | The drawer with the handle | CUTS | +SUBJ<br>+SG<br>+CAN_CUT | +SUBJ<br>+SG | +SG |
| d. | The drawer with the knife | CUTS | +SUBJ<br>+SG<br>+CAN_CUT | +SUBJ<br>+SG | +SG<br>+CAN_CUT |
| e. | The drawer with the handle | CUT | +SUBJ<br>+PL<br>+CAN_CUT | +SUBJ | |
| f. | The drawer with the knives | CUT | +SUBJ<br>+PL<br>+CAN_CUT | +SUBJ | +PL<br>+CAN_CUT |
| g. | The drawer with the knife | CUT | +SUBJ<br>+PL<br>+CAN_CUT | +SUBJ | +CAN_CUT |
| h. | The drawer with the handles | CUT | +SUBJ<br>+PL<br>+CAN_CUT | +SUBJ | +PL |

simulations whether changes in two relevant parameters allow the model to fit the observed pattern qualitatively and quantitatively.

### *5.1. Simulations*

The interACT implementation of the Lewis and Vasishth model introduced two modifications of the original model that are, however, irrelevant for our design (Engelmann et al., 2019).[10] For our simulations, the interACT implementation yields identical results to the original LISP implementation of the Lewis and Vasishth model. Both implementations only allow the use of two retrieval cues. For that reason, we modified interACT to allow the use of three cues needed for the present purposes: structural (indicating whether a noun is in subject position, ±SUBJ), morphosyntactic (±SG, ±PL), and semantic (e.g., ±CAN_CUT). Table 3 shows cue–feature match patterns for all conditions of one example item. The binary semantic features that either fully match or mismatch are likely a simplification of the true state of affairs. For a discussion of how semantic cues can be implemented in a more principled and graded way and how their use can be compared across different models, the reader may refer to Smith and Vasishth (2020).

In the Lewis and Vasishth model, the probability of retrieving a word from memory depends on three parameters:

$$\text{Probability of retrieval} = \frac{1}{1 + e^{\frac{\tau - A}{s}}}$$

We varied two of those parameters: $\tau$ and $s$. Parameter $\tau$ is the *retrieval activation threshold*: the higher the threshold, the lower the probability that some item will be retrieved from memory. If none of the candidates reaches the activation threshold, parsing fails. In ACT-R, the default value of this parameter is 0 (it is $-1.5$ in the Lewis and Vasishth model). We varied it around 0 within the boundaries of $-1.5$ to 1.5 in 13 steps of size 0.25.

Parameter $s$ represents the amount of noise in the system, for example, random fluctuations in activation. These can increase or decrease item activation, which affects the probability of retrieval. The more noise, the less likely it is that the correct item will be retrieved. If noise is close to 0, the transition from low to high probability of retrieval is abrupt (step function), but with greater noise, the transition will follow a smooth sigmoidal function. We varied noise between 0.05 and 0.5 in 10 steps of 0.05 (the default value used in both ACT-R and the Lewis and Vasishth model is 0.2, and in general ACT-R modeling, it is typically varied below 0.5).

We used grid search to systematically vary the two parameters that affect the probability of a retrieval failure: the retrieval activation threshold and the noise parameter. We then identified the set of parameters that most closely reproduced the condition means observed in Experiment 3. Prediction error was quantified in terms of the average mean-squared error across the eight experimental conditions. The simulation was run for 5,000 iterations for each combination of parameter values.

### 5.2. Results

The model predictions generated by the best-fitting set of parameters (retrieval activation threshold: 1, noise: 0.5) are shown in Fig. 4a (gray bars). The model qualitatively predicts all effects we observed: the morphosyntactic and semantic attraction effects both in single and in double subject–verb fit violation conditions, as well as the double attraction effect. The standard error of the model's predictions is below 1%, which means that a difference of several percent between conditions is robust. Quantitatively, the model's predictions lie within the 95% credible intervals for six out of the eight conditions.

Fig. 4b shows how the parameters influence model fit: the retrieval activation threshold affects the fit more than noise, but higher noise values also contribute to a better fit because noise can reduce the activation of the most active item and thus lead to retrieval failures (i.e., correct responses in our task).

To assess whether the predictions of the Lewis and Vasishth model are sufficiently constrained and the model does not predict reverse attraction effects under some parameter configurations, we computed the whole range of model predictions for the attraction effects generated by all possible parameter values, see Fig. 5. The key insight is that Lewis and Vasishth model always predicts correct effect direction (decrease in accuracy due to attraction, blue in the graph) or no effect, but never an incorrect effect direction (increase in accuracy due to attraction, would be red in the graph).

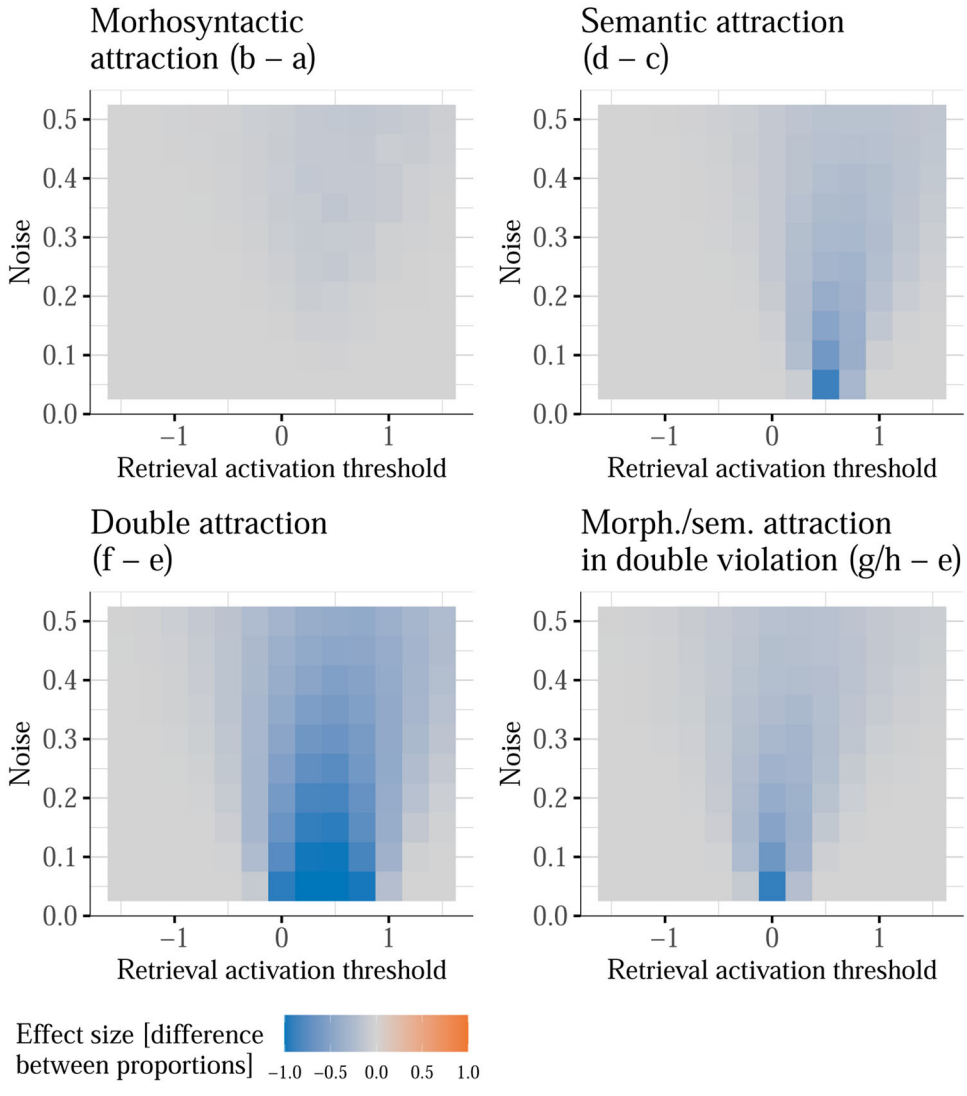


Fig. 5. Attraction effect predicted by Lewis and Vasishth model for acceptability judgments as a function of parameter value. Note that every predicted attraction effect goes in the right direction (blue reflects lower accuracy in attraction conditions). The figure contains only four panels for five attraction effects since from the point of view of the model, semantic and morphosyntactic attraction effects in double verb violation setup are the same, and predictions for conditions (g) and (h) do not differ. Predicted attraction effects are larger for semantic attraction (d–c) than for morphosyntactic attraction (b–a) as in the case of semantic attraction, the attractor matches more features of the verb (see Table 3).

### *5.3. Discussion of simulation results*

We demonstrated that the Lewis and Vasishth model in general predicts the correct direction of all attraction effects in acceptability judgments, and that varying the values of two parameters allowed the model to approximate condition means with a good quantitative fit.

The best-fitting value of the retrieval activation threshold takes on a value (1) that is higher than the default in ACT-R (0) but still plausible. Informally, high retrieval activation threshold represents higher rigor: a structure is built only when the model is confident that it should be built. However, the best-fitting value of noise seems implausible as the estimated value (0.5) was higher than the estimate obtained for participants with aphasia (0.45, Lissón et al., 2021; Mätzig, Vasishth, Engelmann, Caplan, & Burchert, 2018).

## 6. General discussion

Our main goal was to establish whether the well-known agreement attraction effect in sentence comprehension has a counterpart in the semantic domain as was first suggested by reading time data presented by Cunnings and Sturt (2018). In doing so, we also aimed to disambiguate between morphosyntactic theories of agreement attraction, which do not predict semantic attraction effects, and more general sentence processing theories, which do. In three experiments, we replicated the classic morphosyntactic agreement attraction effect,[11] and in all three experiments, we also found highly robust semantic attraction effects[12] that were similar in size to the morphosyntactic attraction effects. Specifically, participants were more likely to accept an unlicensed plural verb as a continuation of a sentence fragment containing a singular subject when another plural noun was present (agreement attraction: "The drawer with the handles open …)." Likewise, participants were also more likely to accept a verb that mismatched the subject semantically as a continuation of the sentence when another noun matched the verb's semantic requirements (semantic attraction: "The drawer with the knife cuts …)." The fact that morphosyntactic and semantic attraction effects were of similar size suggests that both types of errors may be subserved by a common processing substrate. The lack of an interaction between the morphosyntactic and semantic attraction effects is consistent with both a common and with distinct processing substrates.

The Lewis and Vasishth model predicts the observed effects qualitatively. To assess the quantitative fit, we conducted computational simulations with the interACT implementation of the Lewis and Vasishth model (Engelmann et al., 2019). In these simulations, we found that the Lewis and Vasishth model can in principle provide a good quantitative fit to the judgment data. In the following, we briefly discuss the implications of these findings for various theoretical accounts. First, however, we would like to explicitly differentiate our findings from earlier findings that show an involvement of semantic features in agreement attraction.

The aim of our work was to show that semantic features may not only modulate agreement computations, but also give rise to their own attraction effects, that is, lead to attraction-like effects without any involvement of morphosyntactic features, such as number. To illustrate this point and to delineate it from related work, we list studies that investigated the influence of semantics in the context of agreement attraction and explain how these findings differ from ours. The most important difference is that no study tested whether semantic properties of attractor influence the choice of verb stem. Bock and Miller (1991) report that the animacy of the attractor in object relative clauses increases the rate of number attraction errors. While this demonstrates an involvement of semantic features in agreement attraction, it by no means

shows that semantic features can give rise to genuinely semantic attraction effects. Similarly, Thornton and MacDonald (2003) demonstrated that a good thematic fit between the attractor and the verb increased the rate of number attraction errors, but again this did not demonstrate that semantic features can lead to semantic attraction errors. Solomon and Pearlmutter (2004) showed that tight semantic integration between nouns comprising the subject noun phrase increases the rate of number attraction error. Again, no semantic attraction errors where tested or shown. Finally, Schlueter et al. (2019) used semantic match/mismatch to test which particular noun was retrieved when participants were presented with sentences containing number attraction errors, such as "The boy by the trees are really very CHUBBY/GREEN," but the semantic manipulation served only as an instrument to tap into the representation that participants built when processing sentences with agreement attraction errors. Their study did neither test nor demonstrate semantic attraction errors. In sum, all these studies showed the involvement of semantics in the computation of morphosyntactic agreement, but none of them demonstrated purely semantic agreement errors, such as selecting a verb that thematically fits attractor instead of the subject. To our knowledge, the only previous study suggesting that independent semantic attraction may exist is the eye-tracking (during reading) study by Cunnings and Sturt (2018) discussed earlier. Our findings therefore go beyond earlier findings because they demonstrate that the scope of the attraction phenomenon is wider than previously believed, specifically that it extends beyond morphosyntax.

### *6.1. Feature Percolation and Marking and Morphing*

Feature Percolation (Nicol et al., 1997; Vigliocco & Nicol, 1998) and Marking and Morphing (Bock et al., 2001) have both initially been proposed to explain sentence production, not comprehension. However, the fact that agreement attraction effects occur in production and in comprehension and that they largely (but not entirely) pattern together has led researchers to consider these accounts also as explanations for comprehension, where they had some considerable success. For instance, both accounts explain the singular–plural asymmetry (Bock & Cutting, 1992; Bock & Eberhard, 1993; Wagers et al., 2009) and the finding that comprehenders, when asked, tend to report the correct subject but with incorrect number marking (Avetisyan et al., 2020; Paape, Avetisyan, Lago, & Vasishth, 2021; Patson & Husband, 2016; Schlueter et al., 2019).

Given that both these accounts assume the locus of attraction to be in morphosyntactic processing, they make the clear prediction that attraction effects should occur only when morphosyntactic features create the configuration required for attraction to arise. In stark contrast to that prediction, our data show that attraction can also arise without the involvement of morphosyntactic features, and that purely semantic attraction effects not only exist, but that they are as big—which is quite big—as morphosyntactic attraction effects. Feature Percolation and Marking and Morphing therefore do not appear to offer complete explanations for the attraction phenomenon in sentence comprehension.

To incorporate semantic attraction into Feature Percolation and Marking and Morphing, these accounts would need to be either significantly extended by changing some of their core assumptions, or their principles would need to be incorporated into more general models

of attraction mechanism, such as the Lewis and Vasishth model. The latter option is perhaps more attractive as it acknowledges that both types of accounts capture certain aspects of attraction quite well—the singular–plural asymmetry and notional plurality, on the one hand, and semantic attraction effects on the other hand. In sum, we can say that while our data do not support the idea that Feature Percolation and Marking and Morphing offer a complete account of attraction in comprehension, they both still capture important findings and their key ideas therefore remain relevant.

### *6.2. The Lewis and Vasishth model*

The Lewis and Vasishth model was not designed to explain attraction effects. In this model, attraction effects arise from independently motivated ideas about human memory and information processing. The Lewis and Vasishth model has primarily been evaluated on its predictions with respect to morphosyntactic features, and little attention has been given to its predictions involving other kinds of features. Our experiments therefore provide an interesting new perspective on the model. Specifically, our experiments evaluated the surprising prediction that attraction effects are not limited to the morphosyntactic domain, an idea that, despite multiple decades of research on agreement attraction, has, to our knowledge, not received any attention until recently (Cunnings & Sturt, 2018). The Lewis and Vasishth model also predicts that morphosyntactic and semantic attraction effects should have similar or the same size. The fact that both of these predictions were confirmed constitutes strong evidence that this model is capturing key processing principles at the foundations of language comprehension and perhaps language processing and cognition more broadly.

The Lewis and Vasishth model predicts semantic, morphosyntactic, and double attraction effects, and by allowing some parameters to take non-default values, it can closely, though not perfectly, reproduce the observed effects. This suggests that the Lewis and Vasishth model could claim the place of a universal account of attraction phenomena. However, some evidence speaks against that conclusion: First, the value of the noise parameter that provides the best fit to the data is problematic since it has no external justification, and noise levels like those in the optimal model would be more plausible if the participants had language or memory disorders, which was not the case. This shortcoming is alleviated, though, by the fact that lower noise levels still provide a decent fit to the data. See Fig. 4 that shows that with low noise values, such as 0.15, we still obtained a good fit. Nonetheless, this issue poses a challenge for the model and it needs to be addressed in some way. Second, recall that the Lewis and Vasishth model does not cover the full range of findings about agreement attraction effects. The singular–plural asymmetry as well as notional plurality effects lie beyond the scope of the model as currently formulated; though it is in principle possible to import ideas from Feature Percolation and Marking and Morphing to remedy these shortcomings.

The integration of ideas from Feature Percolation and Marking and Morphing into the Lewis and Vasishth model could take many different shapes; a detailed discussion and evaluation is therefore beyond the scope of the present study. It is not clear, for instance, how precisely the percolation of features could be implemented in Lewis and Vasishth model, whether that would even make sense within this model, and how it would affect the models'

predictions for other linguistic structures. Since semantic and morphosyntactic features are believed to have the same status in the model, one would also have to think about the percolation of semantic features, whether that notion makes any sense, and how its predictions could be tested in experiments.

Implementing ideas about markedness of number features might, however, be relatively straightforward. For instance, to account for the singular–plural asymmetry in agreement attraction, it might be sufficient to have only +PL but no corresponding +SG features, as originally proposed by Wagers et al. (2009). As for notional plurality effects, the part of the Marking and Morphing model that accounts for these is already covered by another general sentence processing model, SOSP (Self-Organized Sentence Processing, Smith et al., 2018) in which the effects were successfully modeled by decomposing the abstract concept of notional plurality into several smaller-scale independently motivated semantic features. The Lewis and Vasishth model could perhaps be extended in a similar fashion.

One further challenge for the Lewis and Vasishth model is that it predicts that attraction errors in ill-formed sentences are caused by miscasting of the attractor noun as the subject. That is, the model predicts that the attractor noun is retrieved instead of the subject and subsequently used as the subject leading to an interpretation of the sentence that considerably deviates from the intended interpretation. In contrast to this idea, Schlueter et al. (2019) demonstrated that the attractor noun is used in place of the subject only in a minority of cases. The more common error was that the interpretation used the correct subject but with incorrect number marking, as predicted by encoding accounts of attraction. Although this finding is problematic for the Lewis and Vasishth model, it is in general difficult to establish which noun was retrieved during parsing, as question responses might tap into processes that take place at a later time, for example, posthoc reinterpretation during the question answering period (Bader & Meng, 2018; Meng & Bader, 2021). Either way, this issue remains a conundrum that research on attraction will have to resolve sooner or later.

### *6.3. A reanalysis account of agreement attraction in comprehension*

While our results are largely in line with the predictions of the Lewis and Vasishth model, this is not the only model compatible with our results. Another contender is the reanalysis account proposed by Wagers et al. (2009). Wagers et al. were the first to observe that cue-based parsing as implemented in the Lewis and Vasishth model could potentially explain agreement attraction effects in sentence comprehension (note that this is true only for the processing of ill-formed sentences; in well-formed sentences, the Lewis and Vasishth model predicts a slowdown in number match conditions, see Nicenboim, Vasishth, Engelmann, & Suckow, 2018). However, Wagers et al. also proposed an alternative account that explained their data equally well. The idea is the following: Based on the properties of the subject, the parser predicts a verb with matching properties, and this prediction is always correct, in contrast to the assumptions of Feature Percolation. When the verb in the sentence does not match these expectations, the parser, as a plan B, initiates a retrieval of the subject using the verb's retrieval cues. This is the situation where, according to Wagers et al. (2009), attraction can strike since the attractor matches some of the retrieval cues. Effectively, this may lead the

parser to build an erroneous dependency between the verb and the attractor. Since reanalysis is a prerequisite for attraction, the scope of agreement attraction is rather limited under the reanalysis account: for example, it does not make any predictions about sentence production.

While somewhat ad-hoc, the reanalysis account neatly explains not only Wagers et al.'s data but also Spanish data by Lago et al. (2015). We may therefore ask whether this reanalysis account could also explain semantic attraction. For this account to work, we would have to assume that the parser predicts semantic features of verbs the same way it predicts morphosyntactic features. Further, we would have to assume that repair processes are initiated whenever a verb does not match these semantic predictions. Whether these assumptions are plausible is a matter for debate, but such a model could, in principle, explain our data. Two testable predictions for future work are then (a) that semantic attraction cannot arise when the verb fully matches the subject semantically, and (b) that semantic attraction should not occur in language production.

### *6.4. Production of semantic attraction errors*

The fact that agreement attraction errors largely pattern together in comprehension and production suggests that they may arise from the same underlying principles. If true, we would expect that semantic attraction errors occur not just in comprehension but likewise in production. Specifically, we would expect language users to occasionally produce verbs that match some attractor noun instead of the subject noun in terms of semantic features. This idea might seem bizarre at first blush since such errors could seriously derail a conversation. If true, one might say, why have not we noticed these errors yet? However, the idea that people produce verbs that erroneously match non-subject nouns in morphosyntactic features seems almost equally absurd and unlikely. Nonetheless, it has been attested many times in written and in spoken language (approximately 0.1%–0.5% rate in written corpora according to Stemberger, 1984). Kimball and Aissen (1971), who first observed and discussed the attraction phenomenon, actually thought of the non-standard number agreement that it was a dialect spoken by young people around Boston and did not even consider the possibility that these agreement mismatches may reflect systematic performance errors. But speech errors do happen, even serious ones, and an analysis of spoken or written corpora may very well show that semantic attraction errors occur with some regularity.[13]

Likely scenarios include sentences that are structurally complex and therefore more error-prone in the first place. Further, semantic attraction is perhaps more likely to occur when the subject and the attractor noun are semantically similar. In this case, even a small and potentially benign-looking deviation of the intended verb's semantics may be sufficient to match the attractor's semantics better than the subject's. In these scenarios, semantic attraction errors also may not sound blatantly wrong. Comprehenders may not even notice the errors or they may mentally correct these errors, consciously or not. In sum, it seems at least possible that semantic attraction errors could arise in language production with some regularity even when that may not be our intuition. Only a systematic investigation can tell. This could be a corpus analysis or a clever experiment that attempts to elicit these errors. For the time being, we therefore leave semantic attraction in production as an open question. If future

research convincingly shows that semantic attraction in production is not empirically supported, that would still be an informative finding because it would suggest that attraction in production may be more different from attraction in comprehension than is often assumed. In this case, the reanalysis account by Wagers et al. would become more attractive, because it does not assume mechanisms that we would necessarily expect to be at work in production as well. Any similarities of attraction in production and comprehension would then appear merely superficial.

### *6.5. The single-trial design*

The single-trial design we employed is relatively novel (see, e.g., von der Malsburg, Poppels, & Levy, 2020). Our main motivation for using the procedure was to avoid task adaptation, which has been shown for many experimental paradigms: for reading times (Fine, Jaeger, Farmer, & Qian, 2013), pupil dilations (Demberg & Sayeed, 2016), eye movements in the visual world paradigm (Pregla, Lissón, Vasishth, Burchert, & Stadie, 2021), acceptability judgments (Hammerly et al., 2019; Ness & Meltzer-Asscher, 2021), and speech restoration strategies (Arehalli & Wittenberg, 2021). For a more in-depth discussion of adaptation effects, see Baayen, Vasishth, Kliegl, and Bates (2017) and Arehalli and Wittenberg (2021). In our study, adaptation was a particular concern for two reasons: First, we tested only ill-formed structures, in which case rapid adaptation is expected. Second, in our task, participants could have adopted a superficial processing strategy of reading only the first noun once they noticed that the fit between the first noun and the memory word is all that they needed to pay attention to. This strategy would have seriously threatened the validity of our findings, but the single-trial design made sure that whatever participants learned during the first trial could not inform their performance.

Another strength of the single-trial design was that we obtained data from 4,000+ individuals coming from different continents and from different educational and socioeconomic backgrounds. The data allow much broader generalizations than data coming from a relatively narrow group of, for example, local undergraduates. While the ability to test broad populations is in principle available in all online experiments, it is the single-trial design that leverages this potential to the fullest.

There is one potential downside, however: Single-trial designs are necessarily between-subject designs and these may furnish lower statistical power than otherwise analogous within-subjects designs. However, our prospective power analysis accounted for the between-subjects nature of the design and still indicated acceptable power. This power analysis was also likely too pessimistic since the observed effect, which we replicated five times, was approximately four times bigger than what we assumed in the power analysis. We therefore also conducted an additional post-hoc power analysis that simulated 500 data sets using the effect estimates from Experiment 3. The estimated power in this analysis was 99.6% (95% confidence interval [98.3%, 99.9%]), which suggests that power was not only acceptable but in fact excellent.

It may be surprising that the power was so high given that our experiments used between-subjects designs and given that the number of data points was not higher than in more ordinary

experimental designs (the data points were just spread over a much larger number of participants). We speculate that power was high because we measured the effect upon first exposure of the participants to the stimuli before any adaptation and associated attenuation of effect sizes could kick in.

In sum, we believe that single-trial designs are an interesting alternative to standard repeated-measures designs. A possible limitation of single-trial designs is that they may be less suited for some research questions relating to interindividual differences or when testing populations that are difficult to recruit (patients, children, speakers of certain languages).

### *6.6. Limitations and questions for future research*

Our investigation was limited to the evaluation of sentences with ill-formed subject–verb dependencies. We therefore cannot fully assess the theoretical accounts and the models we considered. Further evaluation on well-formed sentences might provide important insights as the two broad groups of accounts make diverging predictions for such sentences.

While our results are incompatible with the predictions of the faulty encoding accounts, they are compatible with more than only retrieval interference models. One contender that we have already considered is the reanalysis account by Wagers et al. It goes beyond the scope of the present paper to evaluate all the possibilities, but it would be interesting to also formally evaluate the relevant predictions of the Self-Organized Sentence Processing model (SOSP, Smith et al., 2018).

One limitation with regard to our modeling is that neither the Lewis and Vasishth model nor interACT currently account for the human tendency to consider sentences as well-formed by default, as demonstrated by Hammerly et al. (2019). In our simulations, we mapped failed parsing onto rejecting the sentence as ill-formed, but failed parsing might also lead to the acceptance of the continuation due to grammaticality bias in some proportion of cases. Such a modification of the model would affect each condition to a different degree; it is therefore difficult to predict how it would influence simulation outcomes. Investigating the role and influence of response bias therefore remains a task for future research.

A further limitation with regard to modeling concerns the semantic features that we used. Unlike morphosyntactic features, which are well motivated through a large body of research, semantic features are less well understood and potentially more difficult to determine given their often soft nature compared to categorical morphosyntactic features such as number, grammatical gender, and person. While a noun that is marked as singular or masculine is definitely not plural or feminine, such a clear-cut distinction may not be possible for semantic features. The ACT-R framework can account for the fuzzier nature of semantic cues since it allows not only for full, but also for partial matching between retrieval cues and features. The degree of cue–feature match can be set manually for items that are somewhat similar. To mitigate the subjective nature of semantic similarity judgments, Smith and Vasishth (2020) recently proposed an approach for identifying plausible semantic features in a more principled manner, which could be valuable for future investigations on this subject.

## 7. Conclusion

In this study, we presented three experiments ($N_1 = 1,072$, $N_2 = 1,426$, and $N_3 = 2,454$) that demonstrated the existence of a semantic attraction effect in sentence comprehension. This effect closely mirrors the well-known agreement attraction effect. We therefore tentatively suggest that both types of effects may emerge from the same underlying processing substrate and that they may be governed by the same principles. Our findings add to recent reading time evidence presented by Cunnings and Sturt (2018) and help to illuminate the source of these effects.

One consequence of our findings is that attraction may be a much more general phenomenon than is often assumed. Consistent with this notion, the Lewis and Vasishth model naturally predicts that attraction should not be limited to morphosyntax but that it can arise in basically any domain of linguistic description—a rather far-reaching and strong prediction that we partly supported in the present work. It also follows that models assuming that attraction is a primarily morphosyntactic phenomenon may be too narrow in scope. It would be a mistake, however, to consider purely morphosyntactic models obsolete; they still provide better explanations of the circumstances under which morphosyntactic attraction occurs. Based on the present findings and previously published results, we therefore conclude that models of sentence processing, and attraction specifically, need to employ principles and ideas from both schools of thought.

## Acknowledgments

We are grateful to Kate M. Stone and Garrett Smith for their help in creating experimental items, to Garrett Smith and Shravan Vasishth for providing feedback on a previous version of this manuscript. We are especially grateful to Dorothea Pregla who made us aware of a mistake in an earlier version of our simulations. We are also grateful for little Agata and little Otto who are the reason why the publication of this work took much longer than planned.

Open access funding enabled and organized by Projekt DEAL.

WOA Institution: UNIVERSITAET POTSDAM Blended DEAL: Projekt DEAL.

## Notes

1 The +SUBJECT feature is a commonly used simplification, adopted in Jäger, Engelmann, and Vasishth (2017), and elsewhere.

2 Note, though, that some comprehension studies did not find evidence for the singular-plural asymmetry, for example, Häussler (2009) and Acuña-Fariña, Meseguer, and Carreiras (2014).

3 Symmetrical marking of number is perhaps not a strong assumption in the Lewis and Vasishth model and could, in principle, be changed, in which case the model would capture the number asymmetries and similar findings.

4 doi: 10.17605/OSF.IO/P9HS7.

5 LSA is a measure for semantic relatedness based on co-occurrence in text corpora.
6 http://spellout.net/ibexfarm
7 Note that the posterior probability of the parameter being on one side of zero corresponds straightforwardly to the Bayes factor obtained when comparing the hypothesis that the parameter is positive versus negative and when assuming that both hypotheses are equally likely a priori. See Tendeiro and Kiers (2019, 2021) for an explanation.
8 We also analyzed reaction times, but since these were not of primary interest, results are reported in the online Supporting Information.
9 For the last two effects, the percentage values are the rate of correct responses expected if semantic attraction (or double attraction) had the same size as morphosyntactic attraction versus the actual rate of correct responses. For instance, if semantic attraction had the same effect size as morphosyntactic attraction, we would expect 49% correct responses in condition (d) but instead we see 45%, which is non-significantly less.
10 The code of the model is publicly available at https://github.com/felixengelmann/interact/, also available as a Shiny App: https://engelmann.shinyapps.io/inter-act/.
11 Once in Experiment 1, once in Experiment 2 with double violations, and twice in Experiment 3, with single and double violations. One attempted replication with single violations in Experiment 2 produced only a trend in the expected direction.
12 Overall five demonstrations, one in Experiment 1 and two each in Experiments 2 and 3.
13 In written language, semantic attraction errors might be harder to find since they are visually more obvious and therefore more likely to be caught during editing than number agreement errors.
14 We excluded items 1, 2, 12, 19, 20, 21, 22, 23, 25, 26, and 28.

## References

Acuña-Fariña, J. C., Meseguer, E., & Carreiras, M. (2014). Gender and number agreement in comprehension in Spanish. *Lingua*, *143*, 108–128.

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365.

Antón-Méndez, I., Nicol, J., & Garrett, M. F. (2002). The relation between gender and number agreement processing. *Syntax*, *5*(1), 1–25.

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.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, *56*(1), 65–85.

Bader, M., & Meng, M. (1999). Case attraction phenomena in German. Unpublished Manuscript. University of Jena, Jena.

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127.

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

Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, *43*, 83–128.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.

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

Clifton, C., Frazier, L., & Deevy, P. (1999). Feature manipulation in sentence comprehension. *Italian journal of linguistics*, *11*(1), 11–39.

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

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

Deutsch, A., & Dank, M. (2011). Symmetric and asymmetric patterns of attraction errors in producing subject–predicate agreement in Hebrew: An issue of morphological structure. *Language and Cognitive Processes*, *26*(1), 24–46.

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and language*, *36*(2), 147–164.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531–559.

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).

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

Foote, R., & Bock, K. (2012). The role of morphology in subject-verb number agreement: A comparison of Mexican and Dominican Spanish. *Language and Cognitive Processes*, *27*(3), 429–461.

Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition*, *101*(1), 173–216.

Franck, J., Soare, G., Frauenfelder, U. H., & Rizzi, L. (2010). Object interference in subject-verb agreement: The role of intermediate traces of movement. *Journal of Memory and Language*, *62*(2), 166–182.

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4), 371–404.

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.

Harley, H., & Ritter, E. (2002). Person and number in pronouns: A feature-geometric analysis. *Language*, *78*(3), 482–526.

Hartsuiker, R. J., Antón-Méndez, I., & Van Zee, M. (2001). Object attraction in subject-verb agreement construction. *Journal of Memory and Language*, *45*(4), 546–572.

Hartsuiker, R. J., Kolk, H. H., & Huinck, W. J. (1999). Agrammatic production of subject-verb agreement: The effect of conceptual number. *Brain and Language*, *69*(2), 119–160.

Hartsuiker, R. J., Schriefers, H. J., Bock, K., & Kikstra, G. M. (2003). Morphophonological influences on the construction of subject-verb agreement. *Memory & Cognition*, *31*(8), 1316–1326.

Haskell, T. R., & MacDonald, M. C. (2005). Constituent structure and linear order in language production: Evidence from subject-verb agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 891–904.

Häussler, J. (2009). *The emergence of attraction errors during sentence comprehension*. PhD thesis.

Humphreys, K. R., & Bock, K. (2005). Notional number agreement in English. *Psychonomic Bulletin Review*, *12*(4), 689–695.

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.

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

Kimball, J., & Aissen, J. (1971). I think, you think, he think. *Linguistic Inquiry*, *2*(2), 241–246.

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

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.

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

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.

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., Van het Nederend, M. L., Burchert, F., Stadie N., … Vasishth, S. (2021). A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science*, *45*(4), e12956.

Lorimor, H., Bock, K., Zalkind, E., Sheyman, A., & Beard, R. (2008). Agreement and attraction in Russian. *Language and Cognitive Processes*, *23*(6), 769–799.

Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., & Burchert, F. (2018). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, *10*(1), 161–174.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Ranton, FL: CRC Press.

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

Ness, T., & Meltzer-Asscher, A. (2021). Rational adaptation in lexical prediction: The influence of prediction strength. *Frontiers in Psychology*, *12*, 622873.

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34.

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.

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*. Vienna, Austria: R Foundation for Statistical Computing.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, *14*, 611–653.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420.

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

Schlueter, Z., Williams, A., & Lau, E. (2018). Exploring the abstractness of number retrieval cues in the computation of subject-verb agreement in comprehension. *Journal of Memory and Language*, *99*, 74–89.

Slioussar, N., & Malko, A. (2016). Gender agreement attraction in Russian: Production and comprehension evidence. *Frontiers in Psychology*, *7*, 1651.

Slioussar, N., Stetsenko, A., & Matyushkina, T. (2015). Producing case errors in Russian. In *Formal Approaches to Slavic Linguistics: The First New York Meeting* (pp. 363–379).

Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject-verb number agreement. *Cognitive Science*, *42*, 1043–1074.

Smith, G., & Vasishth, S. (2020). A principled approach to feature selection in models of sentence processing. *Cognitive Science*, *44*(12), e12918.

Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, *49*(1), 1–46.

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

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition*, *114*(3), 447–454.

Stemberger, J. P. (1984). Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology*, *1*(4), 281–313.

Sternberg, S. (1998). Discovering mental processing stages: The method of additive factors. In D. Scarborough & S. Sternberg (Eds.), *An invitation to cognitive science: Vol. 4. Methods, models, and conceptual issues* (pp. 703-863). Cambridge, MA: MIT Press.

Tanner, D., & Bulkes, N. Z. (2015). Cues, quantification, and agreement in language comprehension. *Psychonomic Bulletin & Review*, *22*(6), 1753–1763.

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795.

Tendeiro, J. N., & Kiers, H. A. L. (2021). With Bayesian estimation one can get all that Bayes factors offer, and more. https://doi.org/10.31234/osf.io/zbpmy

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), 407–430.

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403.

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., Jäger, L. A., & Nicenboim, B. (2017). Feature overwriting as a finite mixture process: Evidence from comprehension data. *arXiv preprint arXiv:1703.04081*.

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*, *24*(5), 842–848.

Vigliocco, G., Butterworth, B., & Semenza, C. (1995). Constructing subject-verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, *34*(2), 186–215.

Vigliocco, G., & Franck, J. (1999). When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language*, *40*(4), 455–478.

Vigliocco, G., & Franck, J. (2001). When sex affects syntax: Contextual influences in sentence production. *Journal of Memory and Language*, *45*(3), 368–390.
Vigliocco, G., Hartsuiker, R. J., Jarema, G., & Kolk, H. H. (1996). One or more labels on the bottles? Notional concord in Dutch and French. *Language and Cognitive Processes*, *11*(4), 407–442.
Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, *68*(1), B13–B29.
Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology*, *9*, 2.
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.
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.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Dordrecht: Springer.

**Supporting Information**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information

## Appendix A: Plausibility norming for Experiments 1 and 2

LSA scores were analyzed using a linear regression model that included by-item random intercepts and random slopes for plausibility. Factor "plausible" encodes the difference between the two sentence preambles we constructed as plausible and the three preambles constructed as implausible (plausible preambles coded as 1, implausible as −1). The results of the analysis are presented in Table A1.

Table A1
Analysis of semantic similarity scores for Experiments 1 and 2

| Predictor | Estimate | 95%-CrI |
|---|---|---|
| Intercept | 0.13 | 0.10–0.17 |
| Plausible | 0.05 | 0.02–0.08 |

## Appendix B: Plausibility norming for Experiment 3

Ratings were analyzed using an ordinal regression model (Liddell & Kruschke, 2018). Factor "plausible" encodes the difference between the two sentence preambles we constructed as plausible and the three preambles constructed as implausible (plausible preambles coded as

1, implausible as −1). The results of the analysis can be found in Table B1. We excluded four items for which the 95% credible interval for this estimated difference contained 0.

Table B1
Analysis of plausibility ratings for Experiment 3 items

| Predictor | Log-Odds Estimate | 95%-CrI |
|---|---|---|
| Rating 1 | −2.26 | [−2.57, −1.94] |
| Rating 2 | −1.26 | [−1.55, −0.95] |
| Rating 3 | −0.61 | [−0.90, −0.30] |
| Rating 4 | 0.04 | [−0.25, 0.34] |
| Rating 5 | 0.81 | [0.53, 1.12] |
| Rating 6 | 1.90 | [1.60, 2.22] |
| Plausible | 1.58 | [1.27, 1.89] |

## Appendix C: Analysis of items that do not allow for a noun–noun compound interpretation

In this analysis, we included only data from fourteen items[14] that did not allow a noun–noun compound interpretation. We pooled data from Experiments 1 and 2, with resulting $N_{pooled} = 1,426$ (exactly as many as in Experiment 2).

The estimated proportions of correct responses in each condition can be seen in Fig. C1.a and posterior distributions of the parameters in Figs. C1.b and C1.c.

Pooled analysis replicating Experiment 1 (conditions a–d). Accuracy in condition (a) was 82% ($\hat{\beta} = 1.54$, 95%-CrI: [1.08, 2.04]). Accuracy in the semantic and the morphosyntactic baselines (c) and (a) did not differ reliably (82% vs. 72%, $\hat{\beta} = -0.60$, 95%-CrI: [−1.44, 0.28], $P(\beta < 0) = 0.91$). The morphosyntactic attraction effect was present (82% vs. 67%, $\hat{\beta} = -0.82$, 95%-CrI: [−1.32, −0.34], $P(\beta < 0) = 0.99$). The effect of semantic attraction did not differ from the morphosyntactic attraction effect (53% vs. 41%, $\hat{\beta} = -0.47$, 95%-CrI: [−1.17, 0.22], $P(\beta < 0) = 0.91$). When we combined posteriors to estimate the semantic attraction effect directly, we found strong support for the effect being non-zero and in the expected direction ($\hat{\beta} = -1.29$, 95%-CrI: [−1.89, −0.70], $P(\beta < 0) \approx 1$).

Pooled analysis testing the interaction of morphosyntactic and semantic attraction (conditions e–h). Average accuracy across conditions was 83% ($\hat{\beta} = 1.6$, 95%-CrI: [1.2, 2]). Morphosyntactic attraction tended to decrease accuracy (86% vs. 79%, $\hat{\beta} = -0.51$, 95%-CrI: [−1.1, 0.12], $P(\beta < 0) = 0.95$). Semantic attraction reliably decreased accuracy (88% vs. 76%, $\hat{\beta} = -0.87$, 95%-CrI: [−1.4, −0.32], $P(\beta < 0) > 0.99$). There was no evidence for an interaction of morphosyntactic and semantic attraction (83% vs. 83%, $\hat{\beta} = -0.06$, 95%-CrI: [−1, 0.88], $P(\beta < 0) = 0.54$).

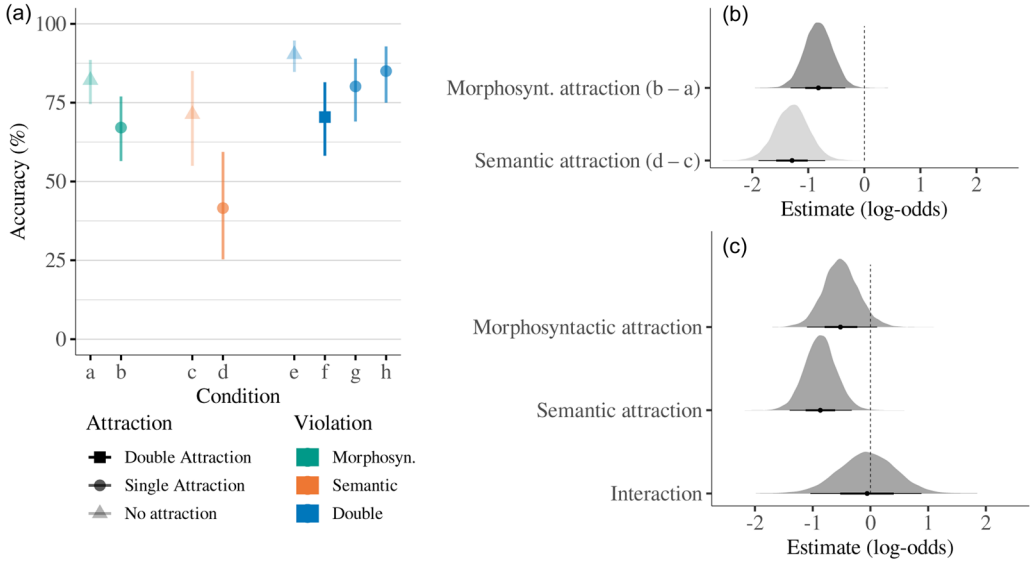


Fig. C1. Analysis of the pooled data from Experiments 1 and 2 (only items that do not allow the noun–noun compound interpretation). Panel A: Estimated condition means with 95% credible intervals. Panel B: Posterior distributions for the model of conditions (a)–(d). The posterior for semantic attraction (light gray) was obtained by combining the posteriors for morphosyntactic attraction and the difference between the semantic and morphosyntactic attraction. Panel C: Posterior distributions for the model of conditions (e)–(h). All parameters are on the log-odds scale. Error bars around the posterior means represent 66% (thick) and 95% (thin) credible intervals.