TOPICS
TOPICS IN COGNITIVE SCIENCE

This article is part of the topic "Best of Papers from the 2024 Cognitive Science Society Conference," Andrea Bender (Topic Editor).

A Working Memory Model of Sentence Processing as Binding Morphemes to Syntactic Positions

Maayan Keshev,[a] ⓘ Mandy Cartner,[b] Aya Meltzer-Asscher,[c] Brian Dillon[d]

[a]*Department of Linguistics, The Hebrew University of Jerusalem*
[b]*Department of Linguistics, Tel-Aviv University*
[c]*Department of Linguistics and Sagol School of Neuroscience, Tel-Aviv University*
[d]*Department of Linguistics, University of Massachusetts Amherst*

Abstract

As they process complex linguistic input, language comprehenders must maintain a mapping between lexical items (e.g., morphemes) and their syntactic position in the sentence. We propose a model of how these morpheme-position bindings are encoded, maintained, and reaccessed in working memory, based on working memory models such as "serial-order-in-a-box" and its SOB-Complex Span version. Like those models, our model of linguistic working memory derives a range of attested memory interference effects from the process of binding items to positions in working memory. We present simulation results capturing similarity-based interference as well as item distortion effects. Our model provides a unified account of these two major classes of interference effects in sentence processing, attributing both types of effects to an associative memory architecture underpinning linguistic computation.

Keywords: Psycholinguistics; Sentence processing; Cognitive modeling; Working memory

---

1. Introduction

To internally represent objects and events, cognitive systems have to maintain an accurate mapping of features to items—for example, that the car ahead is green and the traffic light is red, but not vice versa. Forming such feature-object bindings and maintaining them in working memory is not a trivial task (Treisman, 1996). A similar challenge arguably arises in language processing. Interpreting a sentence requires combining morphemes in an orderly manner—mapping each morpheme to its position in the sentence's structure, a mapping which in turn allows comprehenders to interpret the input in a way consistent with the grammatical structure of a given language.

Here, we propose a model of how this morpheme-structure binding is encoded and maintained in working memory during sentence processing. We adapt a neural net model of item-position mapping in serial recall paradigms (Farrell & Lewandowsky, 2002; Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012) to model the rapid encoding processes that create morpheme-structure bindings. We show that this model can account for a range of attested interference effects in sentence comprehension.

1.1. Linguistic dependencies and interference effects

Morphemes, or morphological features, are the most basic units of syntactic dependencies. For example, in (1), a plural feature is associated with the lexical root morpheme *apprentice*. The noun phrase that all these morphological features comprise in turn occupies the subject position of the sentence, and the agreement on the verb *work* reflects the plural feature of *apprentices*.

(1)   The apprentices work diligently.

The agreement dependency between the verb and its subject is susceptible to interference. For example, an ungrammatical plural verb (e.g., *work* in (2)) could be mistaken as grammatical due to interference from the distractor (e.g., *chefs*), a plural marked noun that is grammatically irrelevant to the subject-verb dependency. This illusion of grammaticality is reflected in increased acceptability ratings, facilitated reading times, and a reduced P600 ERP component (Tanner, Nicol, & Brehm, 2014; Wagers, Lau, & Phillips, 2009). Comprehenders may also experience interference when processing a singular verb in these configurations. That is, *works* in (2) may be judged as unacceptable (Hammerly, Staub & Dillon, 2019) and cause slowed reading times (Laurinavichyute & Malsburg, 2024).

(2)   The apprentice of the chefs work/works diligently.

This pattern of interference is commonly understood to reflect erroneous retrieval of the distractor *chefs* upon reaching the verb (Lewis & Vasishth, 2005; Wagers et al., 2009). However, recent findings from comprehension paradigms show that in the processing of (2), comprehenders may create or maintain a nonveridical representation of the input. Specifically, they appear to misencode a singular subject (e.g., the **target** *apprentice*) as a plural

(e.g., *apprentices*) in configurations like (2), rather than misconstrue the **distractor** *chefs* as the subject (Brehm, Jackson, & Miller, 2021; Paape, Avetisyan, Lago, & Vasishth, 2021).

For example, Keshev, Koesterich, Meltzer-Asscher and Dillon (prep) probed the comprehension of English subject-verb dependencies in sentences like (3), where the past tense verb does not mark the number of the subject. As shown in (3), they tested comprehension with a 4-alternative forced choice task that reveals whether comprehenders represent the subject with the wrong lexical root, the wrong number morpheme, or both. Keshev and colleagues found that, compared to a sentence with a singular distractor, a plural distractor increased the rate of nonveridical target responses (*apprentices*) rather than that of veridical distractor responses (*chefs*).

(3) The apprentice of the chef/chefs worked diligently.
Who worked diligently?
The apprentice / the apprentices / the chef / the chefs

We refer to this type of interpretive error as *item distortion*. These errors bear resemblance to illusory feature conjunctions observed with unattended visual objects (Treisman, 1996): Participants report an interpretation that conjoins the plural morpheme of one item with the lexical root of another. Item distortion errors appear to motivate an alternative account for the illusion in (2), wherein the ungrammatical verb is licensed by erroneous binding of the distractor's plural morpheme (e.g., *-s*) to the subject *apprentice*.

Other types of interference also appear to negatively impact language comprehension. In addition to item distortion errors, semantic similarity between the distractor and the target noun can interfere with access to the target (Smith, Franck, & Tabor, 2021; Van Dyke, 2007). For example, Smith et al. (2021) probed target-distractor confusion in sentences like (4). They found increased error rates when the sentence contained a distractor that is semantically similar to the target, like *kayak* and *canoe*. Importantly, this interference arises even though the features that the nouns share (e.g., the property of being a type of boat) are not probed by the verb (*was damaged*). That is, the verb does not require the subject to be boat-like, and hence is compatible with the semantically distinct distractor *cabin*. This suggests that target-distractor similarity can reduce access to the target *in itself*, independently of distractor-verb compatibility.

(4) The canoe by the cabin/kayak likely was damaged in the heavy storm.
What was damaged in the storm? Canoe/Cabin/Kayak

Gordon, Hendrick and Johnson (2001) and Fedorenko, Gibson, and Rohde (2006) report similar effects in the processing of cleft constructions as in (5). They find that processing the verb inside the cleft (e.g., *saw* in (5)) is slower, and more comprehension errors are observed, when the two nominals are semantically similar: That is, when they are both names, or both professions.

(5) It was the barber / John that the banker / John saw in the parking lot.

To contrast this error pattern with item distortion, we label this type of interference *item confusion*. Whereas *item distortion* occurs when the target and distractor **mismatch**

on a number/gender feature, *item confusion* arises when they **match** in semantic features. Item confusion errors are akin to well-attested similarity-based interference errors in retrieval of word lists, digits, and visual objects (Oberauer, Farrell, Jarrold, & Lewandowsky, 2016).

1.2. Prior models of interference in sentence processing

Interference in sentence processing has overwhelmingly been investigated from the perspective of retrieval processes (Lewis & Vasishth, 2005; Parker, Shvartsman, & Van Dyke, 2017; Wagers & McElree, 2013). For example, the prominent cue-based retrieval model (Lewis & Vasishth, 2005) holds that linguistic input is incrementally encoded into memory as typed feature bundles. These chunks are then reactivated/retrieved when necessary to process subsequent input. This retrieval process is driven by specific retrieval cues. For example, a verb like *thinks* requires an animate, singular subject, and would accordingly initiate a search in content-addressable memory for items with these syntactic and semantic features. Interference on this view arises when the retrieval cues fail to uniquely specify the intended target of retrieval, causing cue overload or misretrieval.

The cue-based retrieval model generally assumes that comprehenders have successfully encoded items in memory such that all morphosyntactic features are bound to the correct chunks in memory (Lewis & Vasishth, 2005). This assumption fails to account for item distortion errors, in which an item appears to be associated with the wrong morphosyntactic features. Moreover, cue-based retrieval links the speed and accuracy of retrieval to the match between cues derived from the retrieval probe and features of the possible retrieval candidates—that is, memory items. Thus, it can account only for a subset of item confusion errors, those where the distractor explicitly matches the retrieval cues. But because interference is cue-driven on this view, this model cannot account for cases of item confusion if the relevant dimension of itemwise similarity does not correspond to a retrieval cue, such as the findings mentioned above by Fedorenko et al. (2006), Gordon et al. (2001), and Smith et al. (2021) (see Logačev & Vasishth, 2011 for additional discussion of non-cue-driven similarity-based interference).

To account for item distortion and non-cue-driven item confusion effects, some researchers have proposed that interference can arise at encoding in addition to at retrieval (Hammerly et al., 2019; Laurinavichyute & Malsburg, 2024; Logačev & Vasishth, 2011; Tanner et al., 2014; Yadav, Smith, Reich, & Vasishth, 2023). We will use the term *encoding interference* to refer to the claim that imperfect memory encodings of the input are an important source of interference effects. The existing literature offers a range of different views, however, on exactly how and why encoding interference may arise. For example, so-called representational models of agreement interference propose that morphosyntactic features can spread from one item to another in memory (Eberhard, Cutting, & Bock, 2005; Hammerly et al., 2019; Yadav et al., 2023). While this feature-passing mechanism has largely been invoked to explain patterns of agreement errors in production, Yadav and colleagues recently showed that hybrid models that invoke cue-based retrieval and this type of encoding distortion provide a superior fit to reading time data than models that only posit retrieval interference (Yadav et al.,

2023). These models provide a potential explanation for item distortion type errors, but do not explicitly capture item confusion errors.

Other models of encoding interference can capture item confusion errors, but fail to capture item distortion errors. Feature Overwriting (e.g., Nairne, 1990) is one such proposal. According to this model, when a new item is encoded into memory, any features it shares with previously encoded items will have some nonzero probability of being deleted or "overwritten." In this way, when a new item shares features with older items, it risks being encoded in memory in a degraded fashion, which impairs any subsequent processing involving that item (Vasishth, Jäger, & Nicenboim, 2017). Another model that captures item confusion errors is Self Organized Sentence Processing (SOSP, Smith, Franck, & Tabor, 2018). This model proposes that sentence processing involves a dynamic process of forming connections between constituent encodings of the sentence. Links between constituents are feature-based and competitive in SOSP: This means that featurewise similarity between multiple elements in memory can inhibit the process of forming a syntactic dependency. Finally, Logačev and Vasishth (2011) proposed a conflicting bindings account of encoding interference. Drawing inspiration from research into the encoding of object files in visual working memory, Logacev and Vasishth proposed that encoding linguistic elements into working memory involves an error-prone process of binding features to chunks. On this view, when a single feature must be bound to multiple distinct items in memory, interference can arise.

While there are a range of models to account for item confusion and item distortion errors separately (or in a hybrid model, e.g., Yadav et al., 2023), to our knowledge, no existing model captures both types of errors in a single architecture. In what follows, we articulate a simple cognitive architecture that aims to derive both types of effects from a single hypothesis about how linguistic input is encoded in working memory. We are inspired by Logacev and Vasishth's proposal that a key challenge in language processing is associating linguistic features with individual items in working memory (Logačev & Vasishth, 2011). However, in our proposal, we suggest that these sources of interference are rooted in how individual items (e.g., words, morphemes, or morphological features) are bound to particular syntactic positions in working memory. We draw an analogy between this process and the process of binding individual items to serial positions in a list—a process that has been studied and modeled extensively. In what follows, we show that a simple distributed neural model that associates item representations with syntactic positions can recreate both item confusion and feature distortion errors.

2. A transient binding model of interference in sentence processing

Our model is a simplified and modified variant of the SOB-CS model of working memory (Oberauer et al., 2012). SOB-CS is based on an earlier model that was developed to explain serial recall data (Farrell & Lewandowsky, 2002, "serial-order-in-a-box"). While the original SOB accounts for well-known serial recall findings such as primacy and recency effects, list length effects, and the distribution of different error types, SOB-CS is designed to account for a range of "benchmark" findings concerning working memory. This includes the findings

that (i) order transposition errors are subject to a clear "locality constraint," such that an item is likely to be erroneously recalled in a list position adjacent or nearly adjacent to its original position; (ii) featural similarity between items in a list can have a detrimental effect on recall; and (iii) there is a greater rate of extra-list intrusions from featurally similar distractor elements (Oberauer et al., 2012).

According to SOB-CS, holding items in working memory involves forming transient associations between items (e.g., words) and position markers (e.g., serial positions in a list). Formally, this is implemented as a two-layer neural network architecture, with one layer representing item information and the other representing position information. Both items and positions are represented as distributed vectors. The vectors encoding item- and position-level information are associated via a fully connected weight matrix $\mathbf{W}$. *Encoding* occurs via a Hebbian update rule that updates the weight matrix W to maintain a new association between a given item $\boldsymbol{v}_i$ and its associated position marker $\boldsymbol{p}_i$:

(6) **Encoding**: $\Delta \mathrm{W} = \eta_e \mathrm{v}_i \mathrm{p}_i{}^T$

The encoding uses the outer product of the item and position vectors as an update to the weight matrix. This update is weighted by an encoding strength parameter η_e. In the SOB-CS model, this parameter takes into account the rate at which information is encoded in memory, the time spent encoding an item, and the item's novelty (see Oberauer et al., 2012). In the simulations reported here, we treat η_e as a free parameter of the model.

Retrieval proceeds by using the vector representing a position marker to reinstate the associated item information encoded in $\mathbf{W}$:

(7) **Retrieval**: $\mathbf{v}_i{}' = \mathbf{W}\mathbf{p}_i$

Where $\mathbf{v}_i{}'$ represents the "retrieved" item information. Crucially, $\mathbf{v}_i{}'$ is not a perfect representation of the original vector encoding $\mathbf{v}_i$: Item information is partially "distorted" by overlapping associations between other positions and items in the weight matrix $\mathbf{W}$. This imperfect retrieved item vector is then compared against all relevant items by computing the cosine similarity $\mathbf{s}_{cos}$ between $\mathbf{v}_i{}'$ and all $\mathbf{v}_j$ in memory.[1] A softmax function is then applied to the resulting similarities to determine the probability of retrieving a given item:

(8) $Pr(\mathbf{v}_j) = \frac{e^{s_{cos}(\mathbf{v}_i{}', \mathbf{v}_j)/\tau}}{\sum_k e^{s_{cos}(\mathbf{v}_i{}', \mathbf{v}_k)/\tau}}$

To apply this model to linguistic structures, we propose that individual morphological formatives—the lexical root and the number agreement morpheme—are encoded independently in the item vector. For purposes of the present simulations, the lexical root is represented by a random vector with 100 dimensions, and the number morpheme is represented by a 20-dimensional vector of 0's (representing SINGULAR), or normalized vector of 1's (representing PLURAL). This encoding assumes that SINGULAR is a default or unmarked state compared to PLURAL. The vectors representing the lexical root and the number morpheme are then concatenated into a single vector representing both morphemes. An item is bound to a 100-dimensional vector $\mathbf{p}_i$ representing position in a hierarchical syntactic structure. Distinct

vectors encode distinct syntactic positions, such as the head and the embedded nominal positions in expressions like *the apprentice of the chefs*.

In all simulations, lexical root vectors and syntactic position vectors were randomly generated as unit vectors constrained to be a given cosine distance from other lexical root or position vectors. This allows us to explore this model's predictions of how interference should be modulated by semantic similarity, that is, similarity between lexical root vectors, and position similarity, that is, similarity between position vectors.

At retrieval, the lexical root and the number morpheme are separately decoded from the reconstituted item vector $\mathbf{v}_i'$. Selecting a certain lexical root for recall involves comparing the units that represent the lexical root in $\mathbf{v}_i'$ ($\mathbf{lex}_i'$) against all lexical roots held in memory, as in (9)a.

(9) a. $Pr(\mathbf{lex}_j) = \frac{\mathbf{s}_{cos}(\mathbf{lex}_i', \mathbf{lex}_j)}{\sum_k \mathbf{s}_{cos}(\mathbf{lex}_i', \mathbf{lex}_k)}$

  b. $Pr(\mathbf{plural}) = \begin{cases} \mathbf{s}_{cos}(\mathbf{num}_i', \mathbf{plural}) & \text{when } \mathbf{s}_{cos} > 0 \\ 0 & \text{otherwise} \end{cases}$

  c. $Pr(\mathbf{v}_j) = Pr(\mathbf{num}_j)Pr(\mathbf{lex}_j)$

The recalled number morpheme is determined by comparing the value of the units that represent the number morpheme in $\mathbf{v}_i'$ ($\mathbf{num}_i'$) against the vector representing PLURAL. The decision criterion in (9) provides a bias for selecting singular: The reconstituted number $\mathbf{num}_i'$ has to be positively correlated with PLURAL, rather than the opposite, to obtain even the smallest probability of plurality. This reflects the assumption that SINGULAR is the default value.

Overall, the fully recalled item in our simulations is sampled from the resulting multinomial distribution of four possible outcomes, crossing the identity of the number feature (SINGULAR vs. PLURAL) and the lexical root (target root or distractor root). Outcome probabilities are set based on the joint distribution of $\mathbf{num}_j$ and each lexical root's $\mathbf{lex}_j$ (product of each number morpheme's probability and each lexical root's probability as in (9)c).

3. Simulations

To simulate the predictions of our model, we generated random vectors for $\mathbf{lex}_i$ and $\mathbf{p}_i$, as well as random starting values for $\mathbf{W}$. In all simulations, the softmax temperature parameter τ was set to 0.1, and the encoding strength parameter η_e was set to 5.[2]

We manipulate cosine similarity between position vectors and between lexical root vectors to determine how position similarity and item similarity impact the results. All results below reflect the average across 100 random runs.

3.1. Item distortion errors

We compare our model against an empirical dataset from Keshev et al. (in prep). As mentioned above, this dataset includes responses to a 4-alternative forced choice task targeting the

(a) Mismatching distractors distort the target's singular feature

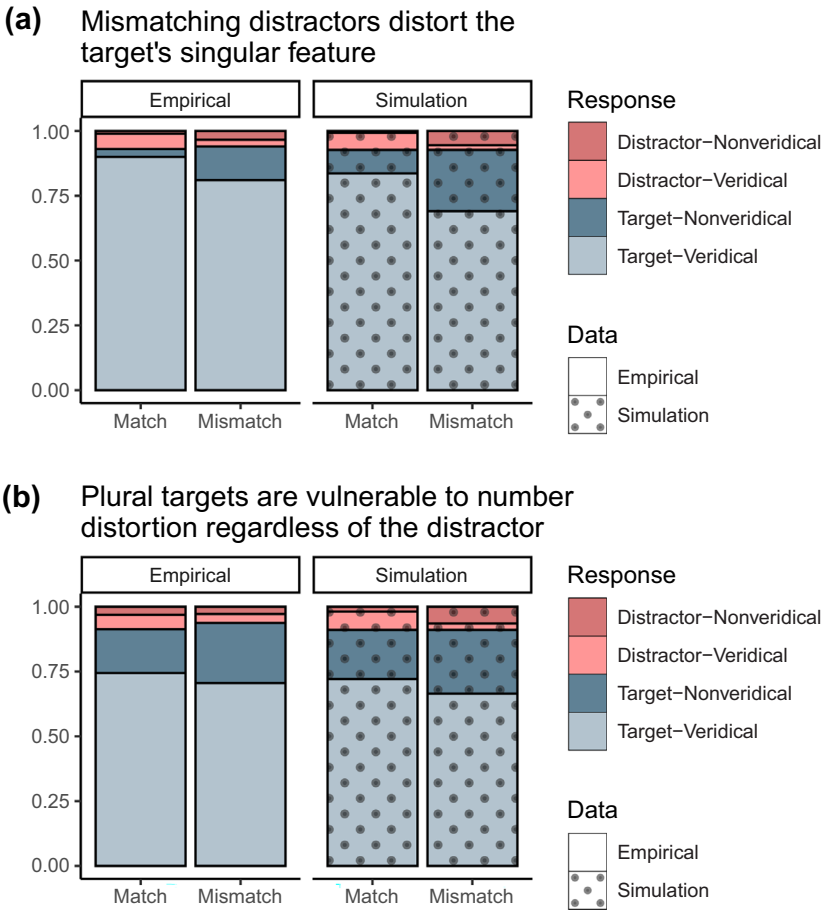**(b)** Plural targets are vulnerable to number distortion regardless of the distractor

Fig. 1. Simulations versus empirical data from Keshev et al. (in prep). Panel A: Results for sentences with target nouns bearing the unmarked (singular) number. Panel B: Results for sentences with target nouns bearing the marked (plural) number. Match/mismatch refers to the match between the target and the distractor's number features. Both simulations use cosine of 0.2 for position vectors (position similarity of 0.2) and for the lexical root vectors (semantic similarity of 0.2).

subject of English sentences, as in (3). This dataset shows that number mismatch between a distractor and a singular target results in distortion of the subject's number, namely, increased rates of nonveridical target choices, that is, *apprentices* in (3) (left panel of Fig. 1a).

Simulations produce a pattern compatible with the one found in the empirical dataset, as depicted in the right panel of Fig. 1a. The simulation shows lower accuracy (namely, lower rate of choosing the veridical target) in the mismatch relative to match conditions, specifically driven by an increase in the rate of responses indicating nonveridical representations of the target noun. Thus, our model can generate item distortion, as observed in this dataset and in previous studies (Brehm et al., 2021; Patson & Husband, 2016; Paape et al., 2021).

Item distortion is known to impact singular (unmarked) subjects more than plural (marked) subjects (Bock & Miller, 1991; Eberhard et al., 2005; Staub, 2009; Wagers et al., 2009). This markedness asymmetry is reflected in the empirical dataset in a diminished contrast between match and mismatch conditions with plural targets (Fig. 1b, left panel), such that the rate of nonveridical target responses is similar across these two conditions, namely, a mismatching distractor does not cause distortion to the same degree as with singular targets. The diminished contrast between match and mismatch conditions for marked (plural) targets has featured prominently in previous models of interference and distortion (Wagers et al., 2009; Eberhard et al., 2005; Smith et al., 2018). However, plural subjects are additionally associated in the empirical dataset with much lower accuracy (i.e., low rates of selecting veridical targets), even in match conditions. This is not an anomaly of the current dataset, as a similar effect can be observed in error rates of preamble repetition in production experiments (Brehm, Cho, Smolensky, & Goldrick, 2022). Yet, this finding has not featured prominently in previous models.

Importantly, our model produces *both* the classical markedness asymmetry, namely, more item distortion with singular compared to plural targets, *and* the low accuracy for plural subjects, as the right panel of Fig. 1b shows. Designating the singular morpheme as an "unmarked" vector (e.g., all 0's) means that it is less disruptive, resulting in attenuation of item distortion with a plural target and singular distractor, compared to a singular target and plural distractor. Additionally, the decoding scheme in (9) yields a singular bias, which leads to lower accuracy with plural compared to singular targets.

3.2. Similarity-based item confusion

Our model assumes separate decoding by morpheme (9): The probability of accessing the target lexical root is independent of the probability of accessing the correct number morpheme and vice versa. Because of this conditional independence, the probability of arriving at each of the four possible representations (veridical/nonveridical target/distractor) is simply the product of $Pr(\mathbf{num}_j)$ and $Pr(\mathbf{lex}_j)$.

Given this assumption, we predict the independence of semantic similarity effects and agreement effects. Since semantic similarity between target and distractor is represented in the lexical root vector, we expect it to increase item confusion (i.e., the rate of choosing the distractor as the root). At the same time, agreement mismatch between target and distractor affects item distortion, that is, the rate of nonveridical representations. The bottom panels of Fig. 2 show the model's prediction of the independence of item confusion and item distortion: Semantic similarity increases distractor choices; agreement mismatch increases nonveridical target choices, and does so to a similar degree irrespective of semantic similarity.

To examine whether these predictions are in line with human performance, we compare our simulation results to empirical data from Laurinavichyute and Malsburg (2024). In this study, the authors manipulated semantic similarity (in terms of animacy) as well as number match between the target and the distractor, as in (10). This publicly available dataset includes results from four high-powered single-trial experiments. Laurinavichute and von der
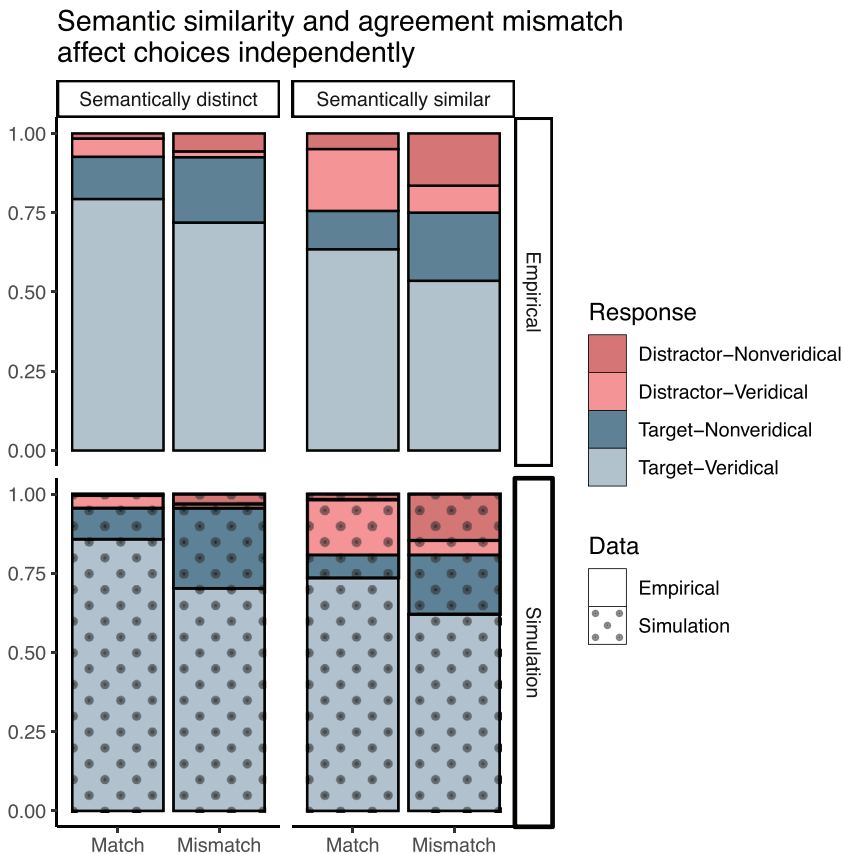
Fig. 2. Simulation results manipulating semantic similarity and empirical data from Laurinavichyute and von der Malsburg (2024). Match/mismatch refers to the match between the target and the distractor's number features. Cosine similarity of the lexical root vectors was 0 for the *semantically distinct* simulation, and 0.5 for the *semantically similar* simulation. Cosine similarity of position vectors was 0.2 for both.

Malsburg's study included, in addition to reading times, a 4-alternative forced choice task probing readers' representation of the subject. The distribution of responses for the comprehension task is depicted in the top panels of Fig. 2. We can see that the pattern of responses in Laurinavichyute and Malsburg's data is compatible with our simulation results, showing independence of semantic similarity and agreement mismatch effects.

(10)   The admirer of the singer(s)/play(s) apparently thinks the show was a big success.
Who considered the show a success?
The admirer / the admirers / the singer / the singers
The admirer / the admirers / the play / the plays

The conclusion that item confusion errors are independent of agreement features might seem at odds with some previous studies. Specifically, increased rates of item confusion
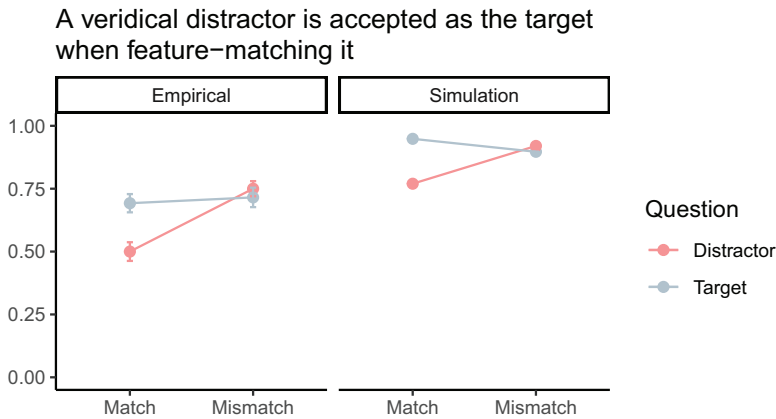
Fig. 3. Simulation results for a yes-no comprehension task and data from Koesterich et al. (2021). Target/distractor refers to whether the question probes the target or the distractor as in (11). Match/mismatch refers to the match between the target and the distractor's gender morpheme. The simulations use cosine of 0.2 for position vectors (position similarity of 0.2) and for the lexical root vectors (semantic similarity of 0.2).

errors have been observed when the target and the distractor match in number (or gender), for subject-verb (Villata, Tabor, & Franck, 2018), anaphor-antecedent (Laurinavichyute et al., 2017), and filler-gap dependencies (Koesterich et al., 2021). Notably, the studies that have detected these types of modulations have predominately detected it via yes/no comprehension questions. For example, Koesterich et al. (2021) tested the comprehension of Hebrew object relative clauses as in (11). They manipulated the match between the distractor's (*manager*) and the target's (*cashier*) grammatical gender (marked on animate nouns in Hebrew). In yes/no comprehension questions, participants were asked either whether the distractor was the object of the embedded verb or whether the target was. Koesterich and colleagues found that readers were less accurate in rejecting matching compared to mismatching distractors as taking the role of the target in the sentence (see left panel of Fig. 3). Thus, item confusion (namely, answering that the distractor is the object of the relative clause verb) was *not* independent of agreement; it increased when the target and distractor matched in gender.

(11) The manager$_{F/M}$ knows the cashier$_F$ that the customers like.
Target Q: Did the customers like the cashier?
Distractor Q: Did the customers like the manager?

We propose that the key difference between these findings (where semantic and agreement effects are not independent) and findings from Laurinavichyute & Malsburg (2024) lies in the comprehension question posed. In Y/N questions (11), the distractor comprehension question probes the probability of reconstructing a *veridical distractor* representation from the retrieved item. Recall that this probability is the product of the probability of recovering the distractor's lexical root from the recovered vector and the probability of interpreting the agreement subspace as the distractor's original morpheme. Importantly, the latter proba-

bility depends on the match between the distractor's and the target's agreement morpheme. This is so since the retrieved vector is generally likely to resemble the target (as the target's position is probed). Therefore, when the distractor shares the agreement morpheme of the target, decoded agreement is more likely to match that of the distractor as well. Thus, the probability of reconstructing a veridical distractor after retrieval is not independent of the agreement manipulation. This is observed not only when the task is a Y/N question probing the veridical distractor representation, as in the current dataset. Rather, it can also be seen in the rates of distractor choices in the previous datasets, in Figs. 1 and 2: agreement match between the target and the distractor increases the rate of selecting a veridical distractor.[3]

To test if this account can predict Koesterich et al.'s (2021) data, we simulate how our model output could be converted to a Yes or No response in this task. We take the probability of reconstructing the veridical target and the veridical distractor representations and add a *yes* bias of 1.5 on the log-odd scale to each. The results are depicted in the right panel of Fig. 3. Although accuracy is higher overall, the simulation produces a pattern compatible with the data from Koesterich et al. (2021), namely, more apparent item confusion when the target and distractor match on gender.

3.3. Order and distance effects

Interference can be either proactive (when the distractor linearly precedes the target) or retroactive (when the distractor follows the target, see Jäger, Engelmann, and Vasishth (2017), for review). Most of the examples so far were of retroactive interference (except (11)). However, in configurations like (12), a plural distractor (*musicians*) that precedes the embedded subject (*reviewer*) readily elicits an illusion of grammaticality at the embedded verb (*praise*) (Wagers et al., 2009). Item confusion may also arise in configurations where a similar distractor (*witness*, in (13)) precedes the target (*attorney*) Van Dyke & McElree, 2011).

(12)   The musicians who the reviewer praise so highly will probably win a Grammy.

(13)   The judge who had declared that the motion/witness was inappropriate realized that the attorney in the case compromised.

In our model, the linear order of encoding items into memory does not affect susceptibility to interference: interference is just as probable for target-distractor and distractor-target orderings. This property is a consequence of the update rule, which encodes new items into memory using simple addition. Since addition is a symmetrical function, the model can capture both proactive and retroactive instances of interference.

Contrarily, structural position is known to modulate interference. Items are less vulnerable to distortion (agreement attraction) when the distractor is structurally distant from the target (Franck, Vigliocco, & Nicol, 2002). In addition, item confusion is affected by the similarity of the target's and the distractor's syntactic position. For example, some studies find that distractors which occupy a subject position interfere more with the processing of other subject-verb
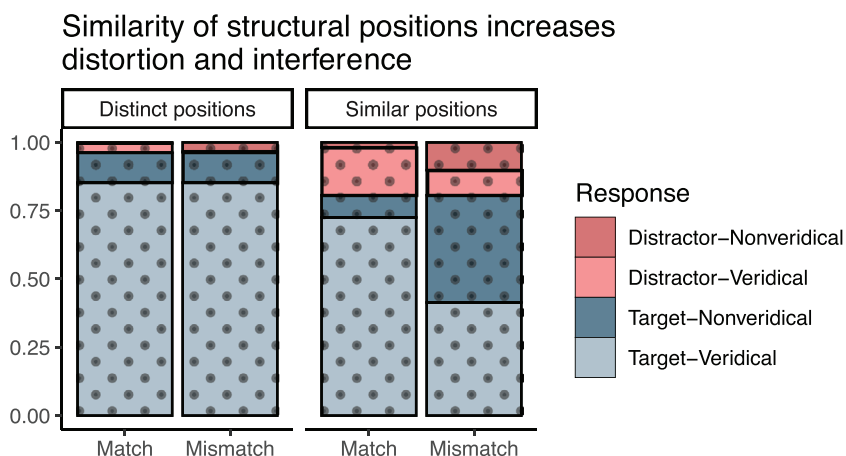
Fig. 4. Simulation results manipulating position similarity. Match/mismatch refers to the match between the target and the distractor's number morphemes. Cosine similarity of the position vectors was 0 for the *distinct positions* simulation, and 0.5 for the *similar positions* simulation. Cosine similarity of lexical root vectors was 0.2 for both.

relations in the sentence (Van Dyke, 2007; Van Dyke & McElree, 2011), see Schoknecht and Vasishth (2023). This interference pattern is sensitive to highly abstract notions of structural similarity (Arnett & Wagers, 2017). Interference might also be selective for intrasentential items (Mertzen, Laurinavichyute, Dillon, Engbert, & Vasishth, 2020).

To test whether our model can capture the effect of position similarity on both types of interference (distortion and confusion), we conducted a simulation manipulating the (dis)similarity in the position vectors to which the target and distractor are bound. We operationalize this similarity as the cosine similarity between the position vectors of each position. Results are shown in Fig. 4. Our model predicts that distractors in positions orthogonal to that of the target (e.g., highly dissimilar positions) do not elicit item distortion due to agreement mismatch. This is reflected in the equal rates of nonveridical responses in the match and mismatch conditions on the left panel of Fig. 4. Increasing the similarity of the positions beyond the similarity used in the previous simulations (to a cosine similarity of 0.5), as in the right panel of Fig. 4, amplifies the match-mismatch contrast in distortion rates. Similarly, orthogonal positions minimize item confusion rates (namely, selecting the distractor instead of the target), but these rates increase with the increase in position similarity between target and distractor.

This property of the model follows from the distributed nature of the position encodings. Each position marker effectively cues not only the item associated with it, but also items associated with partly overlapping position markers. The more the distractor's position vector resembles the target's position vector, the more the distractor will contribute to the vector reconstructed at retrieval (7). Thus, distractors encoded in similar positions (a) are more likely to be confused with the correct item and (b) distort the agreement representation more. Orthogonal position vectors, on the other hand, allow independent encoding of their associated items.

4. Discussion

Interference in sentence comprehension has mostly been researched from the perspective of cue-based retrieval, and takes for granted that comprehenders are able to create unambiguous structure-morpheme mappings (but cf. Futrell, Gibson, & Levy, 2020). This focus neglects a crucial part of working memory's function. We follow (Smolensky, Goldrick, & Mathis, 2014) and propose that, to model the role of working memory in sentence processing, one needs to understand how representations of structure and items are maintained. Our model offers a way of filling this gap from the perspective of encoding and maintaining transient morpheme-position bindings in working memory.

Crucially, we show that one simple mechanism can derive two key types of interference, mismatch-based item distortion and similarly-based item confusion (independent of features of the retrieval trigger). These effects do not receive a full account in the most prominent memory model in sentence processing—cue-based retrieval (Lewis & Vasishth, 2005), and were not previously modeled resulting from a single underlying mechanism (but for a hybrid model, see Yadav et al., 2023). In addition, the proposed model provides insight into previously neglected effects: lower accuracy with plural compared to singular heads, which is predicted by our implementation of markedness; and independence of item confusion errors from agreement errors, predicted by the independent decoding of a lexical root and its agreement morphemes.

A key feature of our model is that item retrieval for dependency completion is based entirely on identification of the relevant structural position, and not on any potentially misleading and context-specific morphological or semantic cues. This reliance on syntactic structure for dependency resolution goes hand-in-hand with the assumptions of theoretical linguistics, wherein syntactic dependencies are constrained by syntactic constraints, such as hierarchical relations and locality. Since structural position is a reliable cue for subject-verb (and many other) dependencies, it seems plausible that comprehenders learn to weigh this cue very highly when possible (Cunnings & Sturt, 2014; Jäger, Mertzen, Van Dyke, & Vasishth, 2020).

Our model also has the potential to capture effects of position similarity. However, our model implemented positional similarity in a very coarse way —by manipulating cosine similarity of randomly generated vectors. Further modeling work is needed to allow principled generation of position vector representations (Cho, Goldrick, & Smolensky, 2020; Smolensky et al., 2014; Smolensky, 1990). This should include position vectors for constituents recursively embedding other items (a key feature of syntactic structure) and a principled conceptualization of position similarity (e.g., operationalized as distance between nodes in a tree, or distributional similarity of constituents, or along the lines explored by Smolensky (1990), Smolensky et al. (2014)). Still, our model provides an interesting testable prediction– - that similarity-based item confusion errors and mismatch-based distortion should both be affected by the same type of syntactic similarity. This is a direct prediction of the model as it binds the lexical root and the agreement morpheme to the same position vector.

Another interesting topic for future research concerns consequences of treating the lexical root as a primitive. We treat morphemes as the basic unit (in the vector's subspaces and in

decoding) and assume distributed vector encodings of the lexical root portion of the item information. This entails that lexical roots should stay intact—they can be confused with one another but no distortion of individual semantic features should arise. However, the links between item information and syntactic position could plausibly come to associate regularly co-occurring sub-lexical semantic features, like animacy, with particular positions, for example, subject position. These learned associations between sub-lexical features and syntactic positions have long been thought to be exploited by comprehenders in parsing linguistic input (MacDonald et al., 1994), suggesting another possible set of results that our proposal may be useful in understanding. Specifically, exploring our model's predictions with more contentful vectors, that is, using contemporary word embedding models, could generate predictions about which semantic features are expected to affect online processing, and how they might come to be associated with syntactic positions in working memory. Beyond this, adopting more theoretically motivated vector encodings may also prove useful in providing grounded models of how morphosyntactic number might be encoded in a vector space. For example, Hao and Linzen (2023) show that, similarly to our model, the large language model BERT maintains a linear encoding of subject number for the purposes of verb inflection. Investigating the diversity of such number encoding strategies in neural language models could prove useful, for example, by helping us to evaluate how robust our model predictions are across different assumptions of how the relevant item information is encoded in a distributed memory.

Lastly, the current model makes broad points of connection with other developments in cognitive science. It emphasizes short/long term memory interactions and connections between semantic memory and an active, goal-directed working memory. It also dovetails with work in deep learning and natural language processing as it highlights the importance of distributed vector representations. In addition, while the comprehension literature has considerably diverged from models of production by focusing on retrieval, our model has the potential of accounting for errors in the two types of processes using the same morpheme to position binding mechanism, which we hope to explore in future research. At the same time, the model highlights the role of structural information as the crucial determinant in the retrieval processes, bridging the memory retrieval tradition in sentence processing and the importance of hierarchical relations in the syntactic literature.

Acknowledgments

Notes

1 Oberauer et al. (2012) use a weighted Euclidean distance metric rather than cosine similarity as their measure of similarity. Our choice of cosine similarity is motivated by its widespread use in Natural Language Processing (Jurafsky & Martin, 2000). Similar results are obtained using Euclidean distance metrics.

2 Parameter values were chosen to set overall reasonable rates of distortion and confusion errors in the first simulation. Shifts to these values did not affect the direction of contrasts between conditions and values were kept consistent in subsequent simulations.

3 Put differently, we suggest that the increased rate of "yes" responses to distractor questions when the distractor and the target match in agreement does not reflect a trade-off with access to the target noun. Instead, acceptance of the distractor trades off with the rate of recovering a nonveridical distractor—a representation which is never probed in those yes/no tasks. We predict that if yes/no questions also feature nonveridical distractors, the combined rate of erroneous yes-responses to both types of distractor should be identical in match and mismatch cases.

References

Arnett, N., & Wagers, M. (2017). Subject encodings and retrieval interference. *Journal of Memory and Language*, *93*, 22–54.

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

Brehm, L., Cho, P. W., Smolensky, P., & Goldrick, M. A. (2022). Pips: A parallel planning model of sentence production. *Cognitive Science*, *46*(2), e13079.

Brehm, L., Jackson, C. N., & Miller, K. L. (2021). Probabilistic online processing of sentence anomalies. *Language, Cognition and Neuroscience*, *36*(8), 959–983.

Cho, P. W., Goldrick, M., & Smolensky, P. (2020). Parallel parsing in a gradient symbolic computation parser. *PsyArXiv*.

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139.

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

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, *9*(1), 59–79.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, *54*(4), 541–553.

Franck, J., Vigliocco, G., & Nicol, J. (2002). Attraction in sentence production: The role of syntactic structure. *Language and Cognitive Processes*, *17*(4), 371–404.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning,Memory, and Cognition*, *27*(6), 1411.

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

Hao, S., & Linzen, T. (2023). Verb conjugation in transformers is determined by linear encodings of subject number. *arXiv preprint arXiv:2310.15151*.

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

Jäger, L., 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.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Keshev, M., Saul, N., Meltzer-Asscher, A., & Dillon, B. (in prep). Feature distortion and memory updating: Experimental and modeling evidence.

Koesterich, N., Keshev, M., Shamai, D., & Meltzer-Asscher, A. (2021). Encoding interference in filler-gap and filler-resumptive dependencies. *CUNY talk*.

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

Laurinavichyute, A., & Malsburg, T. v. d. (2024). *Agreement attraction in grammatical sentences and the role of the task* (Vol. 137).

Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Logačev, P., & Vasishth, S. (2011). Case matching and conflicting bindings interference. In M. Lamers, & P. de Swart (Eds.), *Case, word order and prominence: Interacting cues in language production and comprehension* (pp. 187–216). Springer.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676.

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

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*, 251–269.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity?. *Psychological Bulletin*, *142*(7), 758.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*, 779–819.

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

Parker, D., Shvartsman, M., & Van Dyke, J. A. (2017). The cue-based retrieval theory of sentence comprehension: New findings and new challenges. In L. Escobar, V. Torrens, & T. Parodi (Eds.), *Language processing and disorders* (pp. 121–144).

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

Schoknecht, P., & Vasishth, S. (2023). Do syntactic and semantic similarity lead to interference effects? Evidence from self-paced reading and event-related potentials using German. *PsyArXiv*. Retrieved from https://osf.io/preprints/psyarxiv/cwymg

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

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

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216.

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, *38*(6), 1102–1138.

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

Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, *76*, 195–215.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, *6*(2), 171–178.

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

Van Dyke, J. A., & McElree, B. (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*.

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology*, *9*, 2.

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

Wagers, M., & McElree, B. (2013). Working memory and language processing: Theory, data, and directions for future research. In C. Boeckx, & K. Grohmann (Eds.), *The Cambridge handbook of biolinguistics* (pp. 203–232).

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