A working memory model of sentence processing as binding morphemes to syntactic positions

Maayan Keshev (maayan.keshev@mail.huji.ac.il)
Department of Linguistics, The Hebrew University of Jerusalem
Mount Scopus, Jerusalem 91905, Israel

Mandy Cartner (mandycartner@mail.tau.ac.il)
Department of Linguistics, Tel-Aviv University
Ramat Aviv, Tel Aviv 6997801, Israel

Aya Meltzer-Asscher (ameltzer@tauex.tau.ac.il)
Department of Linguistics and Sagol School of Neuroscience, Tel-Aviv University
Ramat Aviv, Tel Aviv 6997801, Israel

Brian Dillon (bwdillon@umass.edu)
Department of Linguistics, University of Massachusetts Amherst
650 North Pleasant Street, Amherst, MA 01003

Abstract

During sentence processing, comprehenders have to maintain a mapping between lexical items and their position in the sentence (syntactic position). We propose a model of morpheme-position binding in working memory, based on models such as 'serial-order-in-a-box' and its SOB-complex-span version. Like those working memory models, our sentence processing version derives a range of attested memory interference effects from the process of item-position binding. We present simulation results capturing similarity-based interference and item-distortion. These two major classes of interference effects have not received a unified account before, and are not fully captured by cue-based retrieval models.

Keywords: Psycholinguistics; Sentence Processing; Cognitive Modeling; Working Memory

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.

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

Linguistic dependencies and interference effects

Morphemes are the most basic unit of linguistic dependencies. For example, in (1), a plural morpheme is associated with the lexical root *apprentice*. These morphemes are bound to the subject position of the sentence, and the agreement on the verb *work* reflects the plural morpheme 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 (as in (2)) could be mistaken as grammatical due to a interference of a plural distractor (e.g. *chefs*). This illusion of grammatically is reflected in acceptability judgements, fast reading times, and reduced P600 (Wagers, Lau, & Phillips, 2009; Tanner, Nicol, & Brehm, 2014). Comprehenders also occasionally consider a singular verb ungrammatical in those cases. So, *works* in (2) may be judged as unacceptable (Hammerly, Staub, & Dillon, 2019) and incur high reading times (Laurinavichyute & Malsburg, 2024).

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

This pattern of interference could, in principle, reflect erroneous retrieval of the distractor *chefs* upon reaching the verb, or erroneous binding of its plural morpheme *-s* to the subject *apprentice*. Evidence for the latter comes from comprehension paradigms where comprehenders seem to represent a non-veridical item like *apprentices*, (Paape, Avetisyan, Lago, & Vasishth, 2021; Brehm, Jackson, & Miller, 2021).

For example, Keshev et al. (in prep) probed the comprehension of English subject-verb dependencies in sentences like (3) (where the verb does not mark the number of the subject). They probe 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 et al found that a mismatch between a singular target and the plural distractor increased the rate of non-veridical target responses (*apprentices*) rather than that of veridical distractor responses (*chefs*).

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

(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 in visual objects (Treisman, 1996): Participants report an interpretation that conjoins the plural morpheme of one item with the lexical root of another.

A different set of findings shows that semantic similarity between the distractor and the target noun interferes with access to the target (Van Dyke, 2007; Smith, Franck, & Tabor, 2021). For example, Smith et al. (2021) probed target-distractor confusion in sentences like (4). They found increased error rates when the sentence contained a semantically similar distractor like *kayak*. When the distractor is similar to the target, it competes more strongly at retrieval, and can be erroneously activated. This interference arises even though the features that the nouns share are not probed by the verb (*was damaged*). That is, the verb does not require the subject to be boat-like and is compatible with the semantically disctinct distractor *cabin*. This suggests that target-distractor similarity can reduce access to the target 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

To contrast this error pattern with item distortion, we label this outcome *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).

Prior models of interference in sentence processing

The most prominent sentence processing model of memory interference, cue-based retrieval (Lewis & Vasishth, 2005), proposes that memory encodes items as feature bundles. Items are retrieved in response to specific cues in the input. For example, a verb like *think* requires an animate, plural subject. Accordingly, it would initiate a search in content-addressable memory for items with these syntactic and semantic features.

The model, however, does not specify how an accurate encoding of features is achieved. Thus, it does not account for item distortion errors. 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 - i.e. memory items. Thus it can account only for a subset of item confusion errors - depending on distractor-cue compatibility.

Other models of interference in sentence processing also produce only a subset of the above interference types. Representational models of agreement attraction like Marking and Morphing (Eberhard, Cutting, & Bock, 2005) do not capture confusion between similar memory items - they only posit a mechanism for sharing features across items. On the other hand, Feature Overwriting (Nairne, 1990) and Self Organized Sentence Processing (Smith, Franck, & Tabor, 2018) are built to account for similarity-based item confusion and do not incorporate a mechanism for mismatch-based distortion.

In what follows, we articulate a simple cognitive architecture for forming item-to-position bindings in working memory, and show how this architecture derives both item distortion and item confusion interference effects.

A transient binding model of interference in sentence processing

Our model is a 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'). 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 **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 $\boldsymbol{p}_i$:

(5) **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 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 **W**:

(6) **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 encoding $\mathbf{v}_i$: Item information is partially 'distorted' by overlapping associations between other positions and items in the weight matrix **W**. The retrieved item information is then compared against all items in memory 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:

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

To apply this model to linguistic structures, we propose that individual morphemes 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 certain cosine distance from one another. This allows us to explore how interference is affected by semantic similarity, i.e. similarity between lexical root vectors, and position similarity, i.e. 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 (8)a.

(8) 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 (8)b provides a bias for selecting singular: the reconstituted number $\mathbf{num}_i{}'$ has to be positively correlated with PLURAL, rather than not 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 multinomial distribution of four possible outcomes, crossing number (SINGULAR vs. PLURAL) and 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 (8)c).

In all simulations below, 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.

Results

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 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 - i.e. increased rates of non-veridical target choices (left panel of Figure 1A).

Simulations produce a pattern compatible with that in the empirical dataset, as depicted in the right panel of Figure 1A. The simulation shows lower accuracy in the mismatch relative to match conditions, specifically driven by an increase in the rate of responses indicating non-veridical representations of the target noun.

This type of distortion is known to impact singular (unmarked) subjects more that plural (marked) subjects (Bock & Miller, 1991; Wagers et al., 2009). This markedness asymmetry is reflected in a diminished contrast between match and mismatch conditions with plural targets (Figure 1B), such that rate of non-veridical target responses is similar across these two conditions (yet is relatively high). 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 with a much higher base error rate, even in match conditions. This is not an anomaly of the current dataset. A similar effect can be observed in preamble error rate of 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 and the high base rate of error for plural subjects, as Figure 1B shows. Designating the unmarked vector (zeros) to the singular morpheme means that it is less disruptive and the decoding scheme in (8) yields a singular bias. With plural targets, this results in attenuation of the match-mismatch contrast (Figure 1A) and in an overall high

[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 NLP (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.

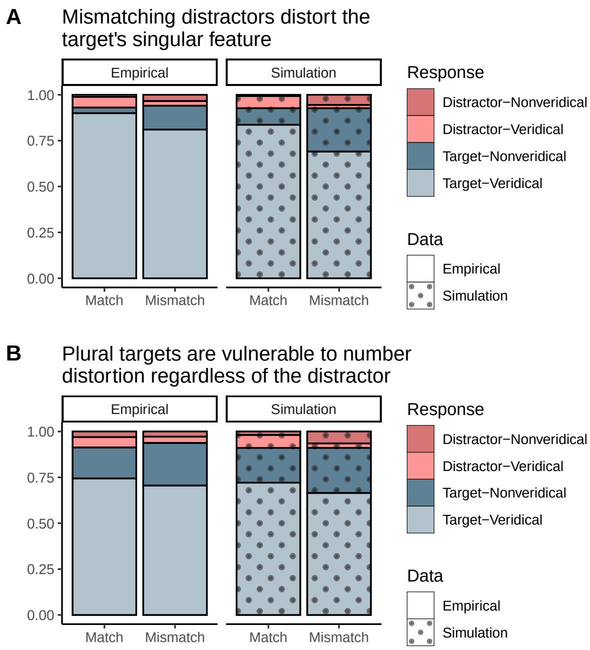

Figure 1: Simulations vs. empirical data from Keshev et al (in prep). Panel A: Results for sentences with target nouns baring the unmarked (singular) number. Panel B: Results for sentences with target nouns baring the marked (plural) number. Match/mismatch refers to the match between the target and the 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).

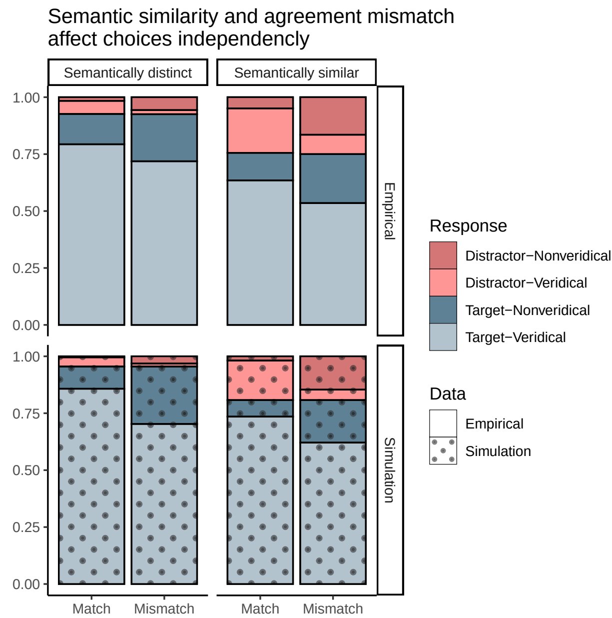

Figure 2: Simulation results manipulating semantic similarity and empirical data from Laurinavichyute and von der Malsburg (2024). Match/mismatch refers to the match between the target and the distractors number features. Cosine similarity of the lexical root vectors was 0 for the *semantically distinct* simulation, and 0.5 for the *semantically similar* simulation. Cosine similarity of position vectors was 0.2 for both.

rate of non-veridical target responses (lower accuracy).

Similarity-based item confusion

Our model assumes separate decoding by morpheme (8). The probability of accessing the distractor lexical root is conditionally independent of the agreement subspace and vice versa. Because of this conditional independence, the probability of arriving at each of the four possible representations (veridical/non-veridical target/distractor) is simply the product of $Pr(\mathbf{num}_j)$ and $Pr(\mathbf{lex}_j)$.

Given this assumption, we expect independence of semantic and agreement effects. Since semantic similarity is represented in the lexical root subspace, we expect it to increase item confusion (i.e. the rate of picking the distractor as the root). On the other hand, agreement-match affects item distortion, i.e. the rate of non-veridical representations. Therefore, item confusion and item distortion should be independent, as can be seen in the bottom panels of Figure 2.

To examine whether these assumptions 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 as well as number match between the target and the distractor, as in (9). This publicly available dataset includes results from four high-powered single-trial experiments. Laurinavichute and von der 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 Figure 2. We can see that the pattern of responses in Laurinavichyute's data is compatible with our simulation results.

(9) 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. Increased rates of item confusion errors have been observed when the target and the distractor match in gender/number, for subject-verb dependencies (Villata, Tabor, & Franck, 2018), anaphors (Laurinavichyute, Jäger, Akinina, Roß, & Dragoy, 2017), and relative clauses (Koesterich, Keshev, Shamai, & Meltzer-Asscher, 2021). The studies that have detected these types of modulations have predominately detected it via yes/no comprehension questions.

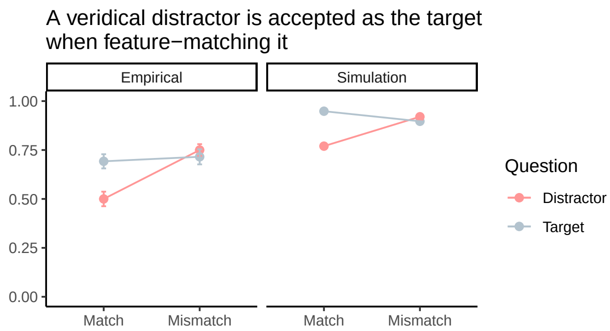

Figure 3: Simulation results for a yes-no comprehension task and data from Koesterich et al (2021). Target/distractor refers to whether the question probes the target or the distractor. as in (10). Match/mismatch refers to the match between the target and the distractor's gender morpheme. The simulation use cosine of 0.2 for position vectors (position similarity of 0.2) and for the lexical root vectors (semantic similarity of 0.2).

For example, Koesterich et al (2021) tested the comprehension of Hebrew object relatives as in (10). 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 et al found that readers were less accurate in rejecting matching compared to mismatching distractors as taking the role of the target in the sentence (see Figure 3).

(10) 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 and findings from Laurinavichyute and Malsburg (2024) lies in the comprehension question posed. In Y/N questions (10), 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 root subspace and the probability of interpreting the agreement subspace as the distractor's original morpheme. The latter probability depends on the match between the distractor's and the target's agreement morpheme.

The retrieved vector is generally likely to resemble the target. Therefore, when the distractor shares the agreement morpheme of the target, decoded agreement is more likely to match that of the distractor as well. Therefore, the probability of reconstructing a veridical distractor after retrieval is not independent of the agreement match manipulation. As an illustration, consider the balance between veridical and non-veridical distractor responses in Figure 1 and 2. This exhibits that a number mismatch between the target and the distractor increases the rate of non-veridical distractor responses.

Overall, we suggest that the increased rate of yes responses to distractor questions 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 non-veridical distractor - a representation which is never probed in those yes/no tasks. We predict that if yes-no questions also feature non-veridical distractors, the combined rate of erroneous yes-responses should be identical in match and mismatch cases.

To test if this can account for Koesterich et al's (2021) data we implement a yes/no version of our model output. 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 Figure 3. The simulation produces a pattern compatible with the data from Koesterich et al (2021).

Order and distance effects

Interference can be either proactive (when the distractor linearly precedes the target) or retroactive (the distractor follows the target, see Jäger et al. (2017), for review). Most of the examples so far are retroactive interference (except (10)). However, in configurations like (11) a plural distractor (*musicians*) that precedes the embedded subject (*reviewer*), readily elicits illusion of grammatically at the embedded verb (*praise*) (Wagers et al., 2009). Similarity-based interference too arises in configurations where a similar distractor (*witness*, in (12)) precedes the target (*attorney*) (Van Dyke & McElree, 2011)).

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

(12) 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 relations in the sentence (Van Dyke, 2007; Van Dyke & McElree, 2011), c.f. Schoknecht and Vasishth (2023). This interference pattern is sensitive to highly abstract notions of structural similarity (Arnett & Wagers, 2017). Interference might also be selective

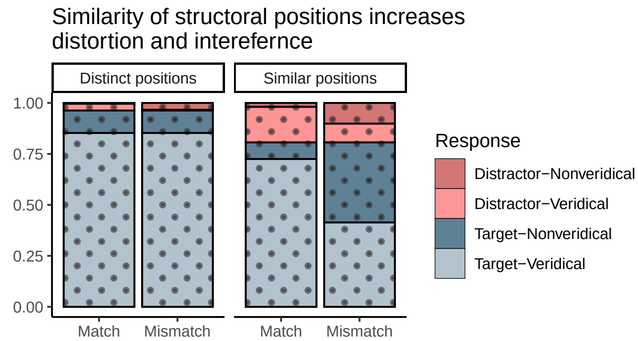

Figure 4: Simulation results manipulating position similarity. Match/mismatch refers to the match between the target and the distractor's number morphemes. Cosine similarity of the position vectors was 0 for the *distinct positions* simulation, and 0.5 for the *similar positions* simulation. Cosine similarity of lexical root vectors was 0.2 for both.

for intra-sentential items (Mertzen, Laurinavichyute, Dillon, Engbert, & Vasishth, 2020).

Our model can capture the modulatory effect on both types of interference. We operationalize the similarity of the structural positions of two different memory items as the cosine similarity between the position vectors (results in Figure 4). Our model predicts that distractors in positions orthogonal to that of the target (e.g. highly dissimilar positions) do not elicit mismatch-based distortion. This is reflected in the equal rates of non-veridical responses on the left panel of Figure 4. Increasing the similarity of the positions beyond the similarity used in the previous simulations (to a cosine similarity of 0.5) amplifies the match-mismatch contrast in distortion rates (i.e. more attraction with increased positional similarity). Similarly, orthogonal positions minimize item-confusion rates, and the rate of choosing the target root approaches chance level with increasing position similarity.

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 (6). 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.

Discussion

Interference in sentence processing 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 et al. (2020)). This focus neglects a crucial part of working memory's function. We follow (Smolensky, Goldrick, & Mathis, 2014) and propose that, to model 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 maintaining transient morpheme-position bindings.

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 description see, Yadav, Smith, Reich, and Vasishth (2023)).

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 (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 encoding at the lexical root subspace. This entails that lexical roots should stay intact - they can be confused with one another but no distortion of individual semantic features should arise.

Lastly, the current model makes broad points of connection with other developments in cognitive science. It emphasizes STM/LTM interactions in working memory and connections between semantic memory and an active, goal directed WM. It also dovetails with work in deep learning and natural language processing as it highlights the importance of distributed vector representations. At the same time, the model highlights the role of structural information as the crucial determinant in the retrieval processes. Thus it bridges the memory retrieval tradition in sentence processing and relational properties of the syntactic literature.

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.

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.

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.

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.

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.

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.

Keshev, M., Koesterich, 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.

Mertzen, D., Laurinavichyute, A., Dillon, B., Engbert, R., & Vasishth, S. (2020). Crosslinguistic evidence against interference from extra-sentential distractors. *PsyArXiv*. Retrieved from `https://psyarxiv.com/t2j8v/`

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.

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.

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.

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.

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.