

# Number feature distortion causes grammaticality illusion in Hindi Sentences

Gaurja Aeron (gaurja@iitgn.ac.in)

Indian Institute of Technology Gandhinagar, Gandhinagar, India

Mukund Choudhary (mukund.choudhary@mbzuai.ac.ae)

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Himanshu Yadav (himanshu@iitk.ac.in)

Indian Institute of Technology Kanpur, 208016 Kanpur, India

## Abstract

How do working memory constraints modulate the real-time sentence comprehension process in humans? A recent proposal is representation distortion, where the representation of sentence material stored in memory probabilistically distorts to a non-veridical representation over time (Yadav et al., 2023; Futrell et al., 2020). The proposal is theoretically important as it challenges the existing theories of sentence comprehension that assume that the representation of sentence material remains intact in memory (Lewis and Vasishth, 2005). However, the proposal lacks empirical testing, and the nature of distortions is ill-understood. We test the representation distortion proposal in Hindi, a verb-final language, using an acceptability judgment study on ungrammatical sentences that can be made grammatical by distorting the number feature on the nouns. We test the representation distortion proposal by implementing four different models that make different assumptions about what kind of distortion occurs: No-distortion model, Insertion error model, Deletion error model, and Swap error model. Data are consistent with the deletion error model, assuming that the plural feature of the nouns can get probabilistically deleted over time. Our work provides a new methodological framework for pinpointing the nature of representation distortions that occur during sentence comprehension.

**Keywords:** representation distortion; sentence comprehension; working memory

## Introduction

To comprehend a sentence, the reader needs to figure out the relationship between arguments (e.g., the subject and the object) and the verb. A key assumption in theories of sentence comprehension is that these dependency relations are identified and built incrementally (Frazier & Fodor, 1978; Frazier, 1987). This process of identifying syntactically related words and linking them together in real time is called **dependency completion**. Dependency completion requires the comprehender to maintain, retrieve, and integrate different chunks of linguistic input in memory. Working memory is argued to play a central role in sentence comprehension (Just & Carpenter, 1992; Lewis, 1996; Gibson, 1998; Lewis & Vasishth, 2005). A well-established working memory-based theory of sentence comprehension is the *cue-based retrieval theory* of Lewis and Vasishth (2005), which maintains that dependency completion is driven by a content-addressable search in memory for syntactically related items. For example, for completing subject-verb dependencies, when the verb is encountered, a search is triggered in memory for the subject noun using

retrieval cues such as *subject* and *plural*. The cue-based retrieval theory enjoys good empirical coverage from a broad range of sentence processing phenomena, including locality, similarity-based interference, and agreement attraction.

A limitation of the cue-based retrieval theory is that it makes a major assumption that when the sentence material is temporarily stored in memory, its representation remains intact (veridical) in memory.

The intact representation assumption has been challenged by a recent theoretical proposal called *representation distortion*: the representation of sentence material stored in memory gets probabilistically distorted to a non-veridical representation over time (Futrell et al., 2020; Yadav et al., 2023; Husain et al., 2025). Such distortions in the representations of the nouns may occur due to feature insertion, deletion, or mis-binding errors (Futrell et al., 2020; Yadav et al., 2023). The representation distortion (RD) proposal has found reasonable empirical support from self-paced reading, eye-tracking, and acceptability judgment studies (Avetisyan et al., 2020; Yadav et al., 2023; Husain et al., 2025).

The RD proposal is theoretically important for two reasons. First, it challenges the conventional view of how working memory processes affect sentence comprehension; the classical models of sentence comprehension hold that the representation of sentence material stored in memory remains intact, i.e., does not undergo any distortion (Lewis & Vasishth, 2005). Second, the proposal finds independent empirical support in visual processing research where feature mis-binding errors are observed when the representation of an object is stored in memory (Bays, 2016); thus, the proposal can contribute to developing a domain-general theory of how working memory limitation drives processing behavior across tasks.

Given the theoretical importance of the representation distortion proposal, we test whether the proposal finds support in Hindi, a verb-final, rich case-marking language. Verb-final languages are important test cases of the working memory-based proposal because they show weak evidence for key working memory effects that are well-attested in English.

We implement the representation distortion proposal as three computational models within the lossy-context surprisal

framework of Futrell et al. (2020). These models differ in their assumptions about what kind of number feature<sup>1</sup> distortions occur in the pre-verbal nouns stored in memory.

1. **The insertion error model:** When pre-verbal nouns are temporarily stored in memory, a plural feature is inserted on the first noun with a probability  $e$ .
2. **The deletion error model:** The plural feature on the first pre-verbal noun is deleted with a probability  $d$ .
3. **The swap error model:** The number features on the first and the second noun are swapped with each other, with a probability  $s$ .

We also implement an *intact representation* proposal that assumes a zero probability of feature distortion when nouns are stored in memory. We plan to test the predictive performance of the above three models against the *intact representation* proposal.

We test the predictions of the above three RD models using data from an acceptability rating experiment in Hindi. The experiment assesses whether pre-verbal nouns undergo representation distortion in processing argument-verb dependencies in Hindi. We record acceptability ratings in ungrammatical sentences that can be made grammatical by minor feature changes (deletion, insertion or swap) on the pre-verbal nouns, to assess what kinds of distortion may occur during comprehension in Hindi. If these ungrammatical sentences are rated as grammatical, it would imply that sentence material stored in memory has probabilistically distorted to a grammatical, non-veridical representation.

We find that the acceptability rating data from our experiments are consistent with the predictions of the deletion error model. The results support *the representation distortion proposal*: when pre-verbal nouns are temporarily stored in memory, they get probabilistically distorted to a non-veridical representation such that the plural feature on the first noun can get deleted with probability  $d$ . Our finding indicates that working memory constraints cause distortions in the memory representation of sentence material temporarily stored in memory.

The paper is organized as follows. First, we present four models of acceptability ratings that make different assumptions about representation distortions. Then, we test the predictions of the models using data from an acceptability rating study on Hindi. Finally, we discuss the findings and conclude.

## Models

We present four models that make different assumptions about what kinds of distortions occur when pre-verbal nouns

<sup>1</sup>In Hindi, the root form of a word is not used in conversations. Generally, and in this study we consider the male, singular, direct case noun/adjective as the default/unmarked form, while the plural counterpart is shown by changing affixes from the singular form and thus considered marked. We define "deletion" and "swap" errors following this assumption.

are stored in memory. The first model is the null hypothesis model that assumes that no feature distortion occurs and sentence material stored in memory stays intact. The other three models (based on representation distortion proposal) assume that when nouns are stored in the memory, the plural number marker on the nouns can get deleted, inserted, or swapped with another noun phrase, influencing the dependency completion process.

## The zero distortion model

The first model is the null-hypothesis model, which assumes no explicit effect of working memory constraints on dependency completion. We take the cue-based retrieval model, which assumes that the dependency completion process is driven by a content-addressable search in the memory through retrieval cues at the verb. It assumes that the pre-verbal input remains intact in the memory during this retrieval process, and any interference is caused due to the presence of other chunks in the sentence which partially match the cues.



Figure 1: Cue-based retrieval process

## Representation distortion models

Working memory limitations may affect the quality of representation of pre-verbal input that is temporarily stored in memory. The representation distortion (RD) models (Futrell et al., 2020; Yadav et al., 2023) assume that the sentence input is probabilistically distorted in memory, such that the comprehender has access only to a non-veridical representation of the actual pre-verbal input. We consider three RD models that differ in their assumptions about what kind of distortions occur: whether the plural number-feature morpheme is inserted, deleted, or swapped probabilistically.

**1. Deletion error model:** The model assumes that the input can distort to a memory representation such that the number feature on the first noun phrase gets deleted with a probability  $d$  (deletion error). The number feature on the second noun phrase does not influence the representation of the number feature on the first noun phrase. For example, when the comprehender stores the input  $NP1.pl\ NP2.pl\ VP.sg$ , in the memory, the first noun phrase can distort to  $NP1.sg$  in some trials.

#### Deletion Error:

\*wo acchhe shikshak chhatr ko padha rha tha.  
 he good.pl teacher student.sg ACC teach PROG PST  
 NP1.pl NP2.sg VP.sg

distorts to:

wo acchha shikshak chhatr ko padha rha tha.  
 he good.sg teacher student.sg ACC teach PROG PST  
 NP1.sg NP2.sg VP.sg

Figure 2: Deletion error

**2. Insertion error Model:** The model assumes that the input can distort to a memory representation such that if the first noun phrase, a plural number feature gets inserted onto it with a probability  $e$  (insertion error). Again, the number feature marking on the second NP does not affect the representation of the first. For example, when the comprehender stores the input *NP1.sg NP2.sg VP.pl*, in some trials, the first noun’s memory representation can become *NP1.pl*.

#### Insertion Error:

\*wo acchha shikshak chhatr ko padha rhe the.  
 he good.sg teacher student.sg ACC teach PROG PST  
 NP1.sg NP2.sg VP.pl

distorts to:

wo acchhe shikshak chhatr ko padha rhe the.  
 he good.pl teacher student.sg ACC teach PROG PST  
 NP1.pl NP2.sg VP.pl

Figure 3: Insertion error

**3. Swap error Model:** The model assumes that the input can distort to a memory representation such that if the first noun phrase has a singular number feature marker and the second noun phrase has a plural marker, the markers will get swapped with probability  $s$  (swap error rate). For example, in the input *NP1.sg NP2.pl VP.sg*, the plural marker on the second noun phrase can get swapped with the singular marker on *NP1* in some trials to produce *NP1.pl NP2.sg VP.sg*.

#### Swap Error:

\*wo acchha shikshak chhatron ko padha rhe the.  
 he good.sg teacher student.pl ACC teach PROG PST  
 NP1.sg NP2.pl VP.pl

distorts to:

wo acchhe shikshak chhatr ko padha rhe the.  
 he good.pl teacher student.pl ACC teach PROG PST  
 NP1.pl NP2.sg VP.pl

Figure 4: Swap error

## Experiment 1: Number-feature distortion in Hindi

To test whether participants undergo representation distortion of pre-verbal nouns, we did an experiment where participants were asked to provide acceptability judgments for Hindi sentences (presented in Devanagari script, see Figure ), displayed on a computer screen in a centered self-paced reading paradigm. After reading the sentences, participants were instructed to rate the sentence on a scale of 0 (not at all acceptable) to 1 (completely acceptable). Some sentences were followed by attention checks (yes/no comprehension questions). Total 79 native Hindi speakers participated in the study, out of which 5 participants were excluded from analysis as they did not meet the criteria of 80% accuracy on attention check comprehension questions. 74 participants were considered for analysis.

The experiment consisted of Hindi sentences across six conditions, with different configurations of (mis)matches in the number feature marking between the subject noun (first NP), the distractor noun (second NP), and the verb phrase. There were 2 grammatical conditions ( (a) and (b), subject noun and verb matched in number feature), and 4 ungrammatical conditions ( (c) to (f), subject noun and verb number feature were mismatched). See Figure 5.

|     | previous class in   | that good teacher                      | that student ACC                     | teach PROG PST                 |
|-----|---------------------|----------------------------------------|--------------------------------------|--------------------------------|
| (a) | pichhli class mein  | wo acchha shikshak<br>NP1.sg (Subject) | uss chhatr ko<br>NP2.sg(Distractor)  | padha rha tha.<br>VP.sg (Verb) |
| (b) | pichhli class mein  | wo acchha shikshak<br>NP1.sg (Subject) | un chhatron ko<br>NP2.pl(Distractor) | padha rha tha.<br>VP.sg (Verb) |
| (c) | *pichhli class mein | wo acchha shikshak<br>NP1.sg (Subject) | un chhatron ko<br>NP2.pl(Distractor) | padha rhe the.<br>VP.pl (Verb) |
| (d) | *pichhli class mein | wo acchha shikshak<br>NP1.sg (Subject) | uss chhatr ko<br>NP2.sg(Distractor)  | padha rhe the.<br>VP.pl (Verb) |
| (e) | *pichhli class mein | wo acchhe shikshak<br>NP1.pl (Subject) | uss chhatr ko<br>NP2.sg(Distractor)  | padha rha tha.<br>VP.sg (Verb) |
| (f) | *pichhli class mein | wo acchhe shikshak<br>NP1.pl (Subject) | un chhatron ko<br>NP2.pl(Distractor) | padha rha tha.<br>VP.sg (Verb) |

Figure 5: Sample Stimuli for the experiment. The sentences mean: “In the previous class, that good teacher(s) was teaching that student(s).”

Each participant rated 40 critical sentences (experimental conditions), and 68 filler sentences. We hypothesized that the representation of the ungrammatical conditions would get distorted in the participants’ memory to a grammatically correct version in some trials, leading them to be rated as acceptable, akin to grammatical conditions. This effect would be observed to be differentially significant for different conditions.

## Predictions

We generate prior predictions from the three representation distortion models for acceptability ratings in the six experimental conditions, two grammatical conditions (a-b) and four

ungrammatical conditions (c-f). The goal is to assess whether a representation distortion model can account for observed acceptability ratings in six conditions from our experiment.

We assume that the acceptability rating (between 0 and 1) on trial  $i$  comes from a Beta distribution with underlying shape parameters  $\alpha$  and  $\beta$ .

$$Rating_i \sim Beta(\alpha_i, \beta_i) \quad (1)$$

where the parameters  $\alpha_i$  and  $\beta_i$  in trial  $i$  are determined by an average rating parameter  $\mu_i$  such that

$$\alpha_i \sim \mu_i \cdot \phi \quad (2)$$

$$\beta_i \sim (1 - \mu_i) \cdot \phi \quad (3)$$

where  $\phi = 1$  and the average rating  $\mu_i$  is a function of three distortion parameters, deletion rate  $d$ , insertion rate  $e$ , and swap error rate  $s$  and the probability of recognizing number agreement/violation  $\theta$ .

$$\mu_i = f(R_i, e, d, s, \theta) \quad (4)$$

Depending on the values of the error rates  $e$ ,  $d$ , and  $s$ , the input representation  $R_i$  distorts to a non-veridical memory representation  $R_j$  with likelihood  $M(R_j|R_i, e, d, s)$ . If the memory representation  $R_j$  is grammatical, it is rated grammatical with probability  $\theta$  and ungrammatical with probability  $(1 - \theta)$ , and vice-versa. The average rating on trial  $i$ ,  $\mu_i$ , is the expected likelihood over memory representations  $M(R_j|R_i)$ .

$$\mu_i = \sum_{j=1}^n M(R_j|R_i, e, d, s) \cdot \theta_{R_j} \quad (5)$$

$$\text{where } \begin{cases} \theta_{R_j} = \theta & \text{if } R_j \text{ is grammatical} \\ \theta_{R_j} = 1 - \theta & \text{if } R_j \text{ is ungrammatical} \end{cases}$$

Using the above assumptions, we simulate prior predictions from each representation distortion model for the six conditions.

**Deletion error model** We make the following prior assumptions about the error rate parameters for the deletion error rate model.

$$d \sim Normal_{lb=0.1, ub=0.5}(0.1, 0.2) \quad (6)$$

$$e = 0 \quad (7)$$

$$s = 0 \quad (8)$$

Figure 6 shows the prior prediction of the deletion error model for the six conditions.

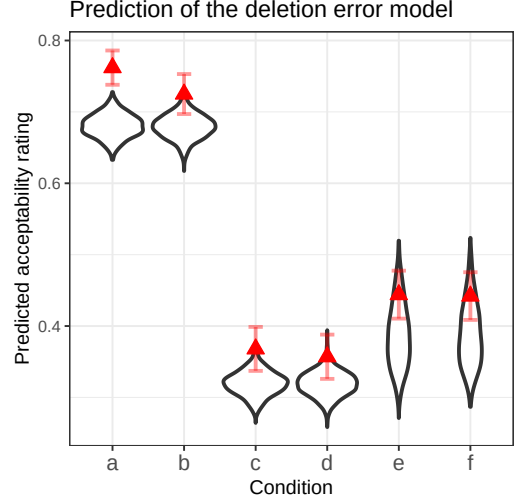


Figure 6: Prior predictions of the deletion error model are shown as violin plots and the observed acceptability rating in each condition is shown as red triangles along with error bars representing 95% confidence intervals.

**Insertion error model** We make the following prior assumptions about the error rate parameters for the insertion error rate model.

$$e \sim Normal_{lb=0.1, ub=0.5}(0.1, 0.2) \quad (9)$$

$$d = 0 \quad (10)$$

$$s = 0 \quad (11)$$

Figure 7 shows the prior prediction of the insertion error model for the six conditions.

**Swap error model** We make the following prior assumptions about the error rate parameters for the swap error rate model.

$$s \sim Normal_{lb=0.1, ub=0.5}(0.1, 0.2) \quad (12)$$

$$d = 0 \quad (13)$$

$$e = 0 \quad (14)$$

Figure 8 shows the prior prediction of the swap error model for the six conditions.

Figure 9 shows the prior prediction of the null hypothesis model that assumes the distortion rate is zero.

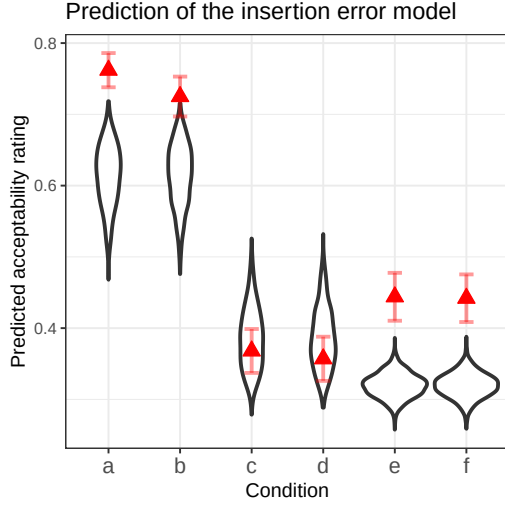


Figure 7: Prior predictions of the insertion error model are shown as violin plots and the observed acceptability rating in each condition is shown as red triangles along with error bars representing 95% confidence intervals.

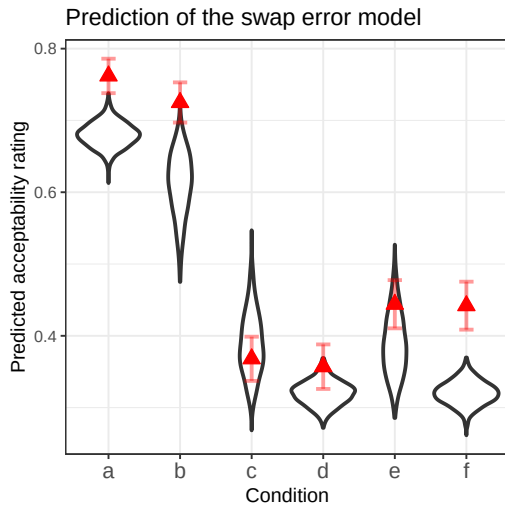


Figure 8: Prior predictions of the swap error model are shown as violin plots and the observed acceptability rating in each condition is shown as red triangles along with error bars representing 95% confidence intervals.

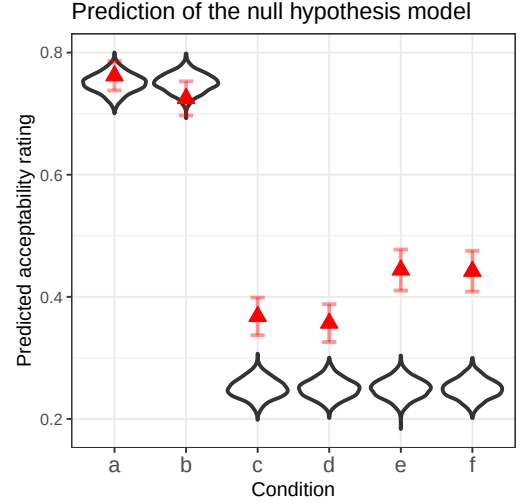


Figure 9: Prior predictions of the null hypothesis model, which assumes zero distortion rate, are shown as violin plots and the observed acceptability rating in each condition is shown as red triangles along with error bars representing 95% confidence intervals.

## Results and Discussion

Table 1: Mean Acceptability Ratings.

| Condition           | Rating           |
|---------------------|------------------|
| (a) Grammatical-1   | 76.2 (se = 1.20) |
| (b) Grammatical-2   | 72.5 (se = 1.41) |
| (c) Ungrammatical-1 | 36.8 (se = 1.54) |
| (d) Ungrammatical-2 | 35.7 (se = 1.55) |
| (e) Ungrammatical-3 | 44.4 (se = 1.68) |
| (f) Ungrammatical-4 | 44.2 (se = 1.67) |

We find three key results:

1. No significant differences were found between the ratings of sentences (a vs. b) and (c vs. d).
2. Conditions (c) and (d) were found to be significantly more acceptable than the conditions (e) and (f), consistent with the key prediction of the deletion error model (see Figure 6)
3. Overall, the representation distortion models perform qualitatively better than the null hypothesis model (Figure 9) at explaining acceptability rating across conditions.
4. The deletion error model shows the best qualitative fit to the acceptability rating data.

Our results imply that the plural number feature marking on the first noun gets probabilistically deleted, impacting the

verb-argument dependency completion process. These findings support the claim that working memory constraints influence sentence comprehension processes through a feature distortion mechanism: when nouns with certain morphological features are stored in the memory, the features may get probabilistically deleted over time.

This study lays a framework for systematic evaluation of the representation distortion proposals (Yadav et al., 2023); the framework used here allows us to pinpoint exactly what kind of feature distortions occur when the sentence material is stored in memory. The other formulations, such as the lossy-context models (Futrell et al., 2020) typically allow a lot of modeler's degree of freedom in specifying the lossy memory encoding function. Our modeling approach allows to systematically test which kind of errors cause the information loss and which types of words or morphemes are lost or changed with time. Consequently, our models generate well-constrained quantitative predictions on a sentence acceptability or a reading task.

Our work is one of the first attempts to empirically estimate the nature of representation distortions and to use these estimates for generating constrained model predictions.

The current work reveals new insights about the working memory process that underlies sentence comprehension: Dependency completion is modulated by probabilistic distortion of the previous linguistic context stored in memory. Our work contributes to understanding how working memory constraints are employed during sentence comprehension.

## Acknowledgments

We thank two anonymous reviewers for their detailed feedback. HY received funding from the Department of Science and Technology (DST) India, under the Cognitive Science Research Initiative (CSRI).

## References

- Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, 112. doi: 10.1016/j.jml.2020.104087
- Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. *Scientific Reports*, 6. doi: 10.1038/srep19203
- Frazier, L. (1987). Syntactic processing: Evidence from dutch. *Natural Language and Linguistic Theory*, 5(4), 519 - 559. doi: 10.1007/BF00138988
- Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6(4), 291 - 325. doi: 10.1016/0010-0277(78)90002-1
- Futrell, R., Gibson, E. A., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3). doi: 10.1111/cogs.12814
- Gibson, E. A. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1 - 76. doi: 10.1016/S0010-0277(98)00034-1
- Husain, S., Apurva, n., Arun, I., & Yadav, H. (2025). The effect of similarity-based interference on bottom-up and top-down processing in verb-final languages: Evidence from hindi. *Journal of Memory and Language*, 143. doi: 10.1016/j.jml.2025.104627
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122 - 149. doi: 10.1037/0033-295X.99.1.122
- Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, 25(1), 93 - 115. doi: 10.1007/BF01708421
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375 - 419. doi: 10.1207/s15516709cog0000.25
- Yadav, H., Smith, G., Reich, S., & Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, 129. doi: 10.1016/j.jml.2022.104400