Evidence for probabilistic feature deletion in sentence comprehension

Himanshu Yadav (himanshu@iitk.ac.in)[1], Gaurja Aeron[2]
[1]Department of Cognitive Science, Indian Institute of Technology Kanpur, [2]Department of Cognitive and Brain Sciences, Indian Institute of Technology Gandhinagar

To comprehend a sentence, the reader must identify syntactically related words and link them together in real time. What are the cognitive processes that underlie real-time structure building in sentence comprehension? A well-established proposal is *cue-based retrieval*: The syntactically related words in a sentence are identified and linked together via a content-addressable search in memory [3]. For example, to integrate a subject and a verb, 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 considerable empirical coverage from multiple sentence processing phenomena including dependency locality and similarity-based interference. A key assumption of the cue-based retrieval theory is the *intact representation assumption*: When the sentence material is temporarily stored in memory, its representation remains intact (veridical) in memory over time. The representation can decay in its accessibility, but it does not change or corrupt in its content.

However, this intact representation assumption has been challenged by a new theoretical proposal, *probabilistic representation distortion*: The representation of sentence material stored in memory gets probabilistically distorted to a non-veridical representation over time, either due to feature deletion, insertion, or misbinding errors [1, 4]. The representation distortion has found reasonable empirical support from reading and acceptability judgment studies [2, 4]. The proposal challenges the conventional view of how working memory constraints determine sentence comprehension.

Given its theoretical importance, we test whether we find support for representation distortion in Hindi, a rich case-marking and verb-final language. We test the predictions of the representation distortion proposal using a new acceptability rating study ($N = 74$) on sentences with a transitive verb and two pre-verbal nouns. We use ungrammatical sentences that can be made grammatical by making minor feature changes, e.g., by deleting the plural marker on the first noun. If an ungrammatical sentence is rated grammatical, it would imply that the sentence material stored in memory has probably been distorted to a grammatical (non-veridical) representation.

We generate predictions from three computational models which make different assumptions about how the number feature of nouns gets distorted over time — (i) *the insertion error model*: A plural feature is inserted with probability e on the first noun stored in memory; (ii) *the deletion error model*: the plural feature on the first pre-verbal noun is deleted with a probability d; and (iii) *the swap error model*: the number features on the first and the second noun are swapped with each other, with a probability s.

We find that the acceptability rating data from our experiment are consistent with the predictions of the deletion error model; the deletion error showed the best qualitative fit to the data compared to other models (see Figure 1). The results support the *representation distortion* proposal: when nouns are temporarily stored in memory, they get probabilistically distorted such that the morphological feature on the first noun gets deleted probabilistically. This study lays a framework for the systematic evaluation of the representation distortion models [1, 4]; the framework allows us to pinpoint exactly what kind of feature distortions occur when the sentence material is stored in memory.

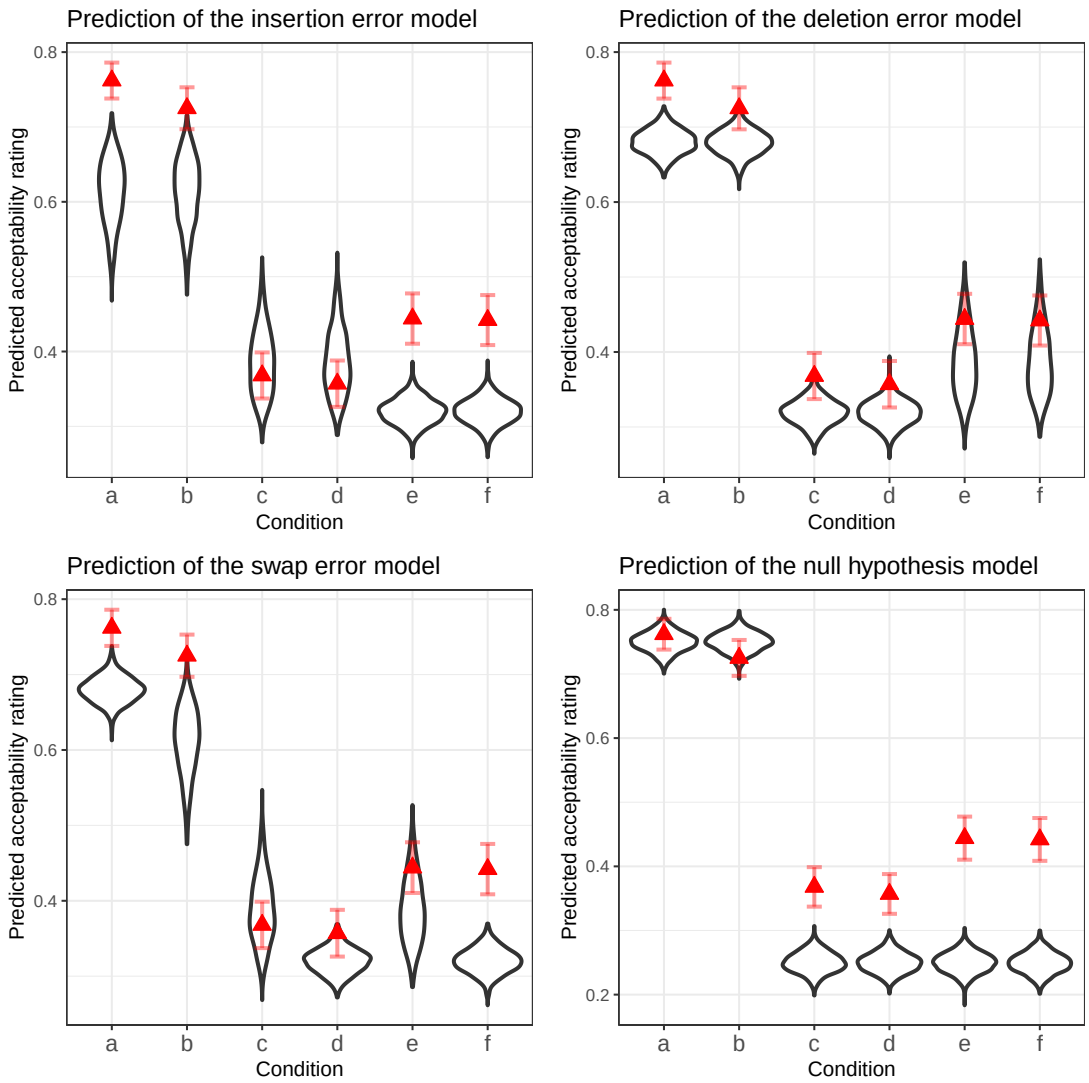

Figure 1: Prior predictions of each model are shown as violin plots, and observed acceptability rating in each condition is shown as red triangles along with error bars (95% confidence interval).

References

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

[2] S. Husain, n. Apurva, I. Arun, and H. Yadav. 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, 2025.

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

[4] H. Yadav, G. Smith, S. Reich, and S. Vasishth. Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, 129, 2023.

Sample stimuli used in the acceptability rating study on Hindi

(a) **Grammatical, Noun 1 - Singular, Noun 2 - Singular**
Last class-DAT that good teacher.sg that student.sg ACC teach.sg PROG Past.

(b) **Grammatical, Noun 1 - Singular, Noun 2 - Plural**
Last class-DAT that good teacher.sg those student-PL ACC teach.sg PROG Past.

(c) **Ungrammatical, Noun 1 - Singular, Noun 2 - Plural**
Last class-DAT that good teacher.sg those student-PL ACC teach.pl PROG Past-PL.

(d) **Ungrammatical, Noun 1 - Singular, Noun 2 - Singular**
Last class-DAT that good teacher.sg that student.sg ACC teach.pl PROG Past-PL.

(e) **Ungrammatical, Noun 1 - Plural, Noun 2 - Singular**
Last class-DAT that good-PL teacher.pl that student.sg ACC teach.sg PROG Past.

(f) **Ungrammatical, , Noun 1 - Plural, Noun 2 - Plural**
Last class-DAT that good-PL teacher.pl those student-PL ACC teach.sg PROG Past.

The above sentences translate to *In the last class, that/those good teacher(s) was/were teaching that/those student(s).*. The exact Hindi sentences used in the experiment are shown in Figure 2.

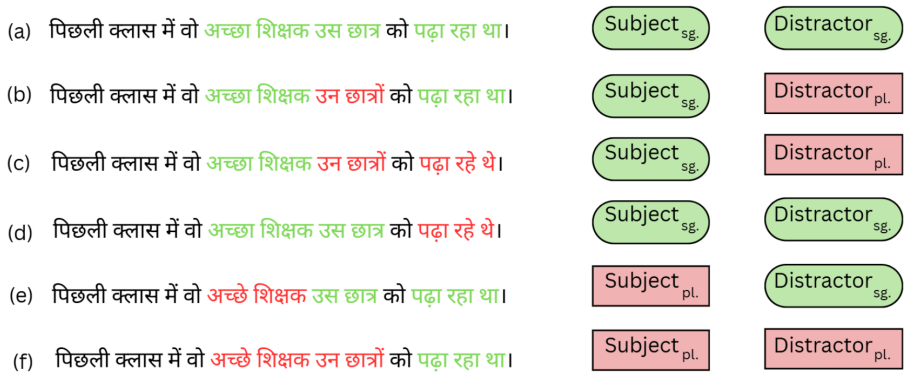

Figure 2: Sample Stimuli for the experiment.

Modeling

We assume that the acceptability rating (between 0 and 1) on trial i comes from a Beta distribution:

$$Rating_i \sim Beta(\alpha_i, \beta_i) \tag{1}$$

where the parameters α_i and β_i in trial i are determined by an average rating parameter μ_i such that

$$\alpha_i = \mu_i \cdot \phi \text{ and } \beta_i = (1 - \mu_i)\phi$$

where the average rating μ_i is a function of three distortion parameters, d, e, and s

$$\mu_i = f(R_i, e, d, s, \theta) \tag{2}$$

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 θ and ungrammatical with probability $(1 - \theta)$, and vice-versa.