Discourse Context Predictability Effects in Hindi Word Order

Sidharth Ranjan
IIT Delhi
sidharth.ranjan03@gmail.com

Marten van Schijndel
Cornell University
mv443@cornell.edu

Sumeet Agarwal
IIT Delhi
sumeet@iitd.ac.in

Rajakrishnan Rajkumar
IISER Bhopal
rajak@iiserb.ac.in

Abstract

We test the hypothesis that discourse predictability influences Hindi syntactic choice. While prior work has shown that a number of factors (e.g., information status, dependency length, and syntactic surprisal) influence Hindi word order preferences, the role of discourse predictability is underexplored in the literature. Inspired by prior work on syntactic priming, we investigate how the words and syntactic structures in a sentence influence the word order of the following sentences. Specifically, we extract sentences from the Hindi-Urdu Treebank corpus (HUTB), permute the preverbal constituents of those sentences, and build a classifier to predict which sentences actually occurred in the corpus against artificially generated distractors. The classifier uses a number of discourse-based features and cognitive features to make its predictions, including dependency length, surprisal, and information status. We find that information status and LSTM-based discourse predictability influence word order choices, especially for non-canonical object-fronted orders. We conclude by situating our results within the broader syntactic priming literature.

1 Introduction

Grammars of natural languages have evolved over time to factor in cognitive pressures related to production (Hawkins, 1994, 2000) and comprehension (Hawkins, 2004, 2014), learnability (Christiansen and Chater, 2008) and communicative efficiency (Jaeger and Tily, 2011; Gibson et al., 2019). In this work, we test the hypothesis that maximization of discourse predictability (quantified using lexical repetition surprisal and adaptive LSTM surprisal) is a significant predictor of Hindi syntactic choice, when controlling for information status, dependency length, and surprisal measures estimated from n-gram, LSTM and incremental constituency parsing models.

Our hypothesis is inspired by a solid body of evidence from studies based on dependency treebanks of typologically diverse languages which show that grammars of languages tend to order words by minimizing dependency length (Liu, 2008; Futrell et al., 2015) and maximizing their trigram predictability (Gildea and Jaeger, 2015). Parallel to this line of work on sentence-level word order, another strand of work has focused on discourse-level estimates of entropy starting from the Constant Entropy Rate hypothesis (CER; Genzel and Charniak, 2002). To overcome the major difficulty of estimating sentence probabilities conditioned on the previous discourse context, Qian and Jaeger (2012) approximated discourse-level entropy using lexical cues from the previous context. In contrast, we leverage modern computational psycholinguistic neural techniques to obtain word and sentence-level estimates of inter-sentential discourse predictability and study the impact of these measures on Hindi word order choices. We conclude that discourse-level priming influences Hindi word order decisions and interpret our findings in the light of the factors outlined by Reitter et al. (2011).

Hindi (Indo-Aryan language; Indo-European language family) has a rich case-marking system and flexible word order, though it mainly follows SOV word order (Kachru, 2006) as exemplified below.

(1) a. amar ujala-ko **yah** *sukravar*-ko daak-se prapt hua
Amar Ujala-ACC **it** *friday*-on post-INST receive be.PST.SG
Amar Ujala received **it** by post on *Friday*.

b. **yah** amar ujala-ko *sukravar*-ko daak-se prapt hua

c. *sukravar*-ko **yah** amar ujala-ko daak-se prapt hua

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10390 - 10406
December 7-11, 2022

To test ordering preferences, we generated meaning-equivalent grammatical variants (Examples 1b and 1c above) of reference sentences (Example 1a) from the Hindi-Urdu Treebank corpus of written text (HUTB; Bhatt et al., 2009) by permuting their preverbal constituent ordering. Subsequently, we used a logistic regression model to distinguish the original reference sentences from the plausible variants based on a set of cognitive predictors. We test whether fine-tuning a neural language model on preceding sentences improves predictions of preverbal Hindi constituent order in later sentences over other cognitive control measures. The motivation for our fine-tuning method is that, during reading, encountering a syntactic structure eases the comprehension of subsequent sentences with similar syntactic structures as attested in a wide variety of languages (Arai et al., 2007; Tooley and Traxler, 2010) including Hindi (Husain and Yadav, 2020). Our cognitive control factors are motivated by recent works which show that Hindi optimizes processing efficiency by minimizing lexical and syntactic surprisal (Ranjan et al., 2019) and dependency length (Ranjan et al., 2022a) at the sentence level.

Our results indicate that discourse predictability is maximized by reference sentences compared with alternative orderings, indicating that discourse predictability influences Hindi word-order preferences. This finding corroborates previous findings of adaptation/priming in comprehension (Fine et al., 2013; Fine and Jaeger, 2016) and production (Gries, 2005; Bock, 1986). Generally, this effect is influenced by lexical priming, but we also find that certain object-fronted constructions prime subsequent object-fronting, providing evidence for self-priming of larger syntactic configurations. With the introduction of neural model surprisal scores, dependency length minimization effects reported to influence Hindi word order choices in previous work (Ranjan et al., 2022a) disappear except in the case of direct object fronting, which we interpret as evidence for the Information Locality Hypothesis (Futrell et al., 2020). Finally, we discuss the implications of our findings for syntactic priming in both comprehension and production.

Our main contribution is that we show the impact of discourse predictability on word order choices using modern computational methods and naturally occurring data (as opposed to carefully controlled stimuli in behavioural experiments). Cross-linguistic evidence is imperative to validate theories of language processing (Jaeger and Norcliffe, 2009), and in this work we extend existing theories of how humans prioritize word order decisions to Hindi.

2 Background

2.1 Surprisal Theory

Surprisal Theory (Hale, 2001; Levy, 2008) posits that comprehenders construct probabilistic interpretations of sentences based on previously encountered structures. Mathematically, the *surprisal* of the k^{th} word, w_k, is defined as the negative log probability of w_k given the preceding context:

$$S_k = -\log P(w_k|w_{1\ldots k-1}) = \log \frac{P(w_1 \ldots w_{k-1})}{P(w_1 \ldots w_k)} \quad (1)$$

These probabilities can be computed either over word sequences or syntactic configurations and reflect the information load (or predictability) of w_k. High surprisal is correlated with longer reading times (Levy, 2008; Demberg and Keller, 2008; Staub, 2015) as well as longer spontaneous spoken word durations (Demberg et al., 2012; Dammalapati et al., 2021). Lexical predictability estimated using *n*-gram language models is one of the strongest determinants of word-order preferences in both English (Rajkumar et al., 2016) and Hindi (Ranjan et al., 2022a, 2019; Jain et al., 2018).

2.2 Dependency Locality Theory

Dependency locality theory (Gibson, 2000) has been shown to be effective at predicting the comprehension difficulty of a sequence, with shorter dependencies generally being easier to process than longer ones (Temperley, 2007; Futrell et al., 2015; Liu et al., 2017, cf. Demberg and Keller, 2008).

3 Data and Models

Our dataset comprises 1996 reference sentences containing well-defined subject and object constituents from the HUTB[1] corpus of dependency trees (Bhatt et al., 2009). The HUTB corpus, which belongs to the newswire domain and contains written text in a natural discourse context,

[1] https://verbs.colorado.edu/hindiurdu/

is a human-annotated, multi-representational, and multi-layered treebank. The dependency trees here assumes Panini's grammatical model where each sentence is represented as a series of *modifier-modified* elements (Bharati et al., 2002; Sangal et al., 1995). Each tree in the HUTB corpus denotes words in the sentence with nodes such that head words (modified) are linked to the dependent words (modifier) via labelled links denoting the grammatical relationship between word pairs.

For each reference sentence in the HUTB corpus, we created artificial variants by permuting the preverbal constituents whose heads were linked to the root node in the dependency tree. Inspired by grammar rules proposed in the NLG literature (Rajkumar and White, 2014), ungrammatical variants were automatically filtered out by detecting dependency relation sequences not attested in the original HUTB corpus. After filtering, we had 72833 variant sentences for our classification task. Figure 1 in Appendix A displays the dependency tree for Example sentence 1a and explains our variant generation procedure in more detail.

To determine whether the original word order (i.e. the reference sentence) is preferred to the permuted word orders (i.e. the variant sentences), we conducted a targeted human evaluation via forced-choice rating task and collected sentence judgments from 12 Hindi native speakers for 167 randomly selected reference-variant pairs in our data set. Participants were first shown the preceding sentence, and then they were asked to select the best continuation between either the reference or the variant. We found that 89.92% of the reference sentences which originally appeared in the HUTB corpus were also preferred by native speakers compared to the artificially generated grammatical variants expressing similar meaning (Further details are provided in Appendix G). Therefore, in our analyses we treat the HUTB reference sentences as human-preferred gold orderings compared with other possible automatically-generated constituent orderings.

3.1 Models

We set up a binary classification task to separate the original HUTB reference sentences from the variants using the cognitive metrics described in Section 2. To alleviate the data imbalance between the two classes (1996 references vs 72833 variants), we transformed our data set using the approach described in Joachims (2002). This technique converts a binary classification problem into a pairwise ranking task by training the classifier on the difference of the feature vectors of each reference and its corresponding variants (see Equations 2 and 3). Equation 2 displays the objective of a standard binary classifier, where the classifier must learn a feature weight (w) such that the dot product of w with the reference feature vector ($\phi(reference)$) is greater than the dot product of w with the variant feature vector ($\phi(variant)$). This objective can be rewritten as equation 3 such that the dot product of w with the difference of the feature vectors is greater than zero.

$$w \cdot \phi(reference) > w \cdot \phi(variant) \quad (2)$$

$$w \cdot (\phi(reference) - \phi(variant)) > 0 \quad (3)$$

Every variant sentence in our dataset was paired with its corresponding reference sentence with order balanced across these pairings (e.g., Example 1 would yield (1a,1b) and (1c,1a)). Thereafter, their feature vectors were subtracted (e.g., 1a-1b and 1c-1a), and binary labels were assigned to each transformed data point. *Reference-Variant* pairs were coded as "1" and *Variant-Reference* pairs were coded as "0". The alternate pair ordering thus re-balanced our previously severely imbalanced classification task. Table 5 in Appendix D illustrates the original and transformed values of the independent variables.

For each reference sentence, our objective was to model the possible syntactic choices entertained by the speaker. In each instance, the author chose to generate the reference order over the variant, implicitly demonstrating an order preference. If the cognitive factors in our study influenced that decision, a logistic regression model should be able to use those factors to predict which syntactic choice was ultimately chosen by the author. Using the transformed features dataset labelled with 1 (denoting a preference for the reference order) and 0 (denoting a preference for the variant order), we trained a logistic regression model to predict each reference sentence (see Equation 4). We report our classification results using 10-fold cross-validation. The regression results are reported on the entire transformed test data for the respective

experiments. All experiments were done with the Generalized Linear Model (GLM) package in *R*.

$$choice \sim \begin{cases} \delta \text{ dependency length} + \\ \delta \text{ trigram surp} + \delta \text{ pcfg surp} + \\ \delta \text{ IS score} + \delta \text{ lexical repetition surp} + \\ \delta \text{ lstm surp} + \delta \text{ adaptive lstm surp} \end{cases} \quad (4)$$

Here *choice* is encoded by the binary dependent variable as discussed above (1: reference preference and 0: variant preference). To obtain sentence-level surprisal measures, we summed word-level surprisal of all the words in each sentence. The values for independent variables were calculated as follows.

1. **Dependency length**: We computed a sentence-level dependency length measure by summing the head-dependent distances (measured as the number of intervening words) in the HUTB reference and variant dependency trees.

2. **Trigram surprisal**: For each word in a sentence, we estimated its local predictability using a 3-gram language model (LM) trained on the written section of the EMILLE Hindi Corpus (Baker et al., 2002), which consists of 1 million mixed genre sentences, using the SRILM toolkit (Stolcke, 2002) with Good-Turing discounting.

3. **PCFG surprisal**: The syntactic predictability of each word in a sentence was estimated using the Berkeley latent-variable PCFG parser[2] (Petrov et al., 2006). 12000 phrase structure trees were created to train the parser by converting Bhatt et al.'s HUTB dependency trees into constituency trees using the approach described in Yadav et al. (2017). Sentence level log-likelihood of each test sentence was estimated by training a PCFG LM on four folds of the phrase structure trees and then testing on a fifth held-out fold.

4. **Information status (IS) score**: We automatically annotated whether each sentence exhibited *given-new* ordering. The subject and object constituents in a sentence were assigned a *Given* tag if its head was a pronoun or any content word within it was mentioned in the preceding sentence. All other phrases were tagged as *New*. For each sentence, IS score was computed as follows: a) Given-New order = +1 b) New-Given order = -1 c) Given-Given and New-New = 0. For an illustration of givenness coding, see Example 3 in Appendix A and the description in Appendix B.

5. **Lexical repetition surprisal**: For each word in a sentence, we accounted for lexical priming by interpolating a 3-gram language model with a unigram cache LM based on the history of words ($H = 100$) containing only the preceding sentence. We used the original implementation provided in the SRILM toolkit with a default interpolation weight parameter ($\mu = 0.05$; see Equations 5 and 6) based on the approach described by Kuhn and De Mori (1990). The idea is to keep a count of recently occurring words in the sentence history and then boost their probability within the trigram language model. Words that have occurred recently in the text are likely to re-occur[3] in subsequent sentences (Kuhn and De Mori, 1990; Clarkson and Robinson, 1997).

$$P(w_k|w_1, w_2, ...w_{k-1}) = \mu\, P_{cache}(w_k|w_1, w_2, ...w_{k-1}) + (1-\mu)\, P_{trigram}(w_k|w_{k-2}, w_{k-1}) \quad (5)$$

$$P_{cache}(w_k|w_{k-H}, w_{k-H+1}, ..w_{k-1}) = \frac{w_k\ counts_{(cache)}}{H} \quad (6)$$

6. **LSTM surprisal**: We estimated the predictability for each word according to the entire sentence prefix using a long short-term memory language model (LSTM; Hochreiter and Schmidhuber, 1997) trained on the 1 million written sentences from the EMILLE Hindi corpus (Baker et al., 2002). We used the LSTM implementation provided in the Neural Complexity toolkit (van Schijndel and Linzen, 2018) with default hyper-parameter settings to estimate surprisal using the neural context within each sentence. The exact parameters are as follows: 2 LSTM layers with 200 hidden units each, 200-dimensional word embeddings, 20 units each of learning rate,

[2] 5-fold cross-validated parser training and testing F1-score metrics were 90.82% and 84.95%, respectively.

[3] Out of 13274 sentences present in HUTB, 71.20% sentences contained at least one content word previously mentioned in the preceding sentence (Jain et al., 2018).

Learning Rate	0	0.002	0.02	0.2	**2**	20	200
Perplexity	103.29	98.79	87.78	66.64	**56.86**	117.91	$\sim 10^9$

Table 1: Learning rate influence on adaptive LSTM validation perplexity ($N = 13274$ sentences; the initial non-adaptive model uses a learning rate of 0)

and training epoch with early stopping. Rest other parameters were set to default setting.[4]

7. **Discourse LSTM surprisal**: We estimated the discourse predictability of each word in the sentence using the ADAPT function of the neural complexity toolkit. van Schijndel and Linzen (2018) proposed a simple way to continuously adapt a neural LM to each successive test sentence, and found that adaptive surprisal predicts human reading times significantly better than non-adaptive surprisal. Their method takes a pre-trained LSTM LM, and, after generating surprisals for a test sentence, the parameters of the LM get updated based on the cross-entropy loss for that sentence. After that, the revised LM weights are used to predict the next test sentence. This continuous fine-tuning approach effectively modulates a sentence-level LSTM through discourse priming. In our work, for each test sentence, we used our base LSTM LM and adapted it to the immediately preceding context sentence and then used it to generate (discourse-sensitive) surprisal values for the desired sentence. We used an adaptive learning rate of 2 as it minimized the perplexity of the validation data set (see Table 1).[5]

4 Experiments and Results

We tested the hypothesis that discourse predictability (estimated from adaptive LSTM and lexical repetition surprisal) influences constituent ordering in Hindi over other baseline cognitive controls, including dependency length, information status and trigram and non-adaptive surprisal. The adaptive LSTM surprisal had a high correlation with all other surprisal features and a low correlation with dependency length and information status score (see Figure 2 in Appendix C). We report the results of regression and prediction experiments on the full data set as well as on subsets of the data consisting of two types of non-canonical constructions.

[4] https://github.com/vansky/neural-complexity#model-parameters

[5] Interestingly, van Schijndel and Linzen (2018) found that an adaptive learning rate of 2 minimized validation perplexity in English as well, though we leave further investigation of this to future work.

Predictor	$\hat{\beta}$	$\hat{\sigma}$	t
intercept	1.50	0.001	1496.47
trigram surprisal	-0.08	0.005	-14.53
dependency length	0.02	0.001	15.55
pcfg surprisal	-0.07	0.002	-39.46
IS score	0.01	0.001	11.32
lex-rept surprisal	-0.03	0.005	-5.31
lstm surprisal	-0.14	0.016	-9.26
adaptive lstm surprisal	-0.13	0.016	-8.18

Table 2: Regression model on full data set ($N = 72833$; all significant predictors denoted by |t|>2)

4.1 Regression Analysis

Our regression results over the entire data set (Table 2) show that all of our measures are significant predictors for the task of classifying reference and variant sentences. The negative regression coefficients for our surprisal metrics (including adaptive LSTM surprisal) indicate that surprisal is consistently lower in the reference sentences compared with the competing variants. And adding adaptive discourse LSTM surprisal into a model containing all other predictors significantly improved the fit of our regression model (χ^2 = 66.81; p $<$ 0.001). Thus these results support our core hypothesis that word order choices seem to maximize discourse predictability compared with possible alternative productions. The positive regression coefficient for information status (IS) score indicates that reference sentences adhere to *given-new* ordering. Similarly, adding IS score into a model containing all other predictors significantly improved the fit of our regression model (χ^2 = 127.94; p $<$ 0.001). However, the positive regression coefficient of dependency length suggests that reference sentences exhibit *longer* dependency lengths compared to their variant counterparts, violating locality considerations. Thus dependency length might be in conflict with (and/or overridden by) other factors like discourse priming or information locality (see Section 6 for more discussion of this idea).

We also examined the contribution of each predictor on two non-canonical syntactic configura-

tions, *direct object (DO) fronted* and *indirect object (IO) fronted* constructions, which have been studied extensively in the sentence comprehension literature. Prior work has shown that salient objects tend to occur early in the sentence, thus leading to fronting (Wierzba and Fanselow, 2020; Kaiser and Trueswell, 2004). In the specific context of Hindi, Vasishth (2004) examined the role of locality effects in processing these non-canonical word orders in salient as well as non-salient contexts. He showed that the increased distance to the verb in DO-fronted sentences leads to high self-paced reading times at the inner-most verb as compared to its canonical counterpart in both salient and non-salient conditions. However, in IO-fronted constructions, he found that salient contexts alleviated the processing difficulty which was caused by increased distance. Based on these findings, we predict that **adaptive surprisal should be more effective in IO-fronted than DO-fronted constructions**.

To test this hypothesis, we isolated reference sentences where the direct object precedes the subject (for a DO-fronted test set) and reference sentences where the indirect object precedes the subject (for an IO-fronted test set) along with their context sentences. We compared both sets to paired variants that exhibited canonical order (i.e. where the subject preceded both objects). Tables 3a and 3b present regression results for DO- and IO-fronted constructions respectively. These subsets constitute a very small fraction of our dataset due to the infrequency of these constructions in Hindi. The regression coefficient for adaptive LSTM surprisal was significantly negative for both subsets, indicating that the non-canonical structures are more common in the context of similarly non-canonical structures. This pattern is more robust for IO-fronted reference sentences (χ^2 = 90.90; p $<$ 0.001) than for DO-fronted reference sentences (χ^2 = 4.03; p = 0.04), validating our proposed prediction about these constructions. Coming to the efficacy of IS scores over these two non-canonical constructions, *givenness* is effective in case of DO-fronted reference sentences only (χ^2 = 49.06; p $<$ 0.001). Furthermore, in contrast to the IO-fronted subset, the regression coefficient for dependency length in DO-fronted items is significantly negative suggesting that locality considerations are limited to constructions involving a high dependency length difference between reference and variants,[6] a similar finding to that reported in Ranjan et al. (2022a) on the same task.

4.2 Prediction Accuracy

While the previous section explored how predictors contribute to Hindi ordering preferences across all of the data in aggregate, in this section we frame our model as a classification task on held-out data to determine how many sentences are affected by each predictor. This enables us to examine the relative performance of different predictors in identifying Hindi reference sentences amidst artificially generated grammatical variants and to conduct more detailed error analysis of our results. We used 10-fold cross-validation to evaluate model classification accuracy, i.e. the percentage of data points where a model correctly predicted the referent sentence over a paired variant, for different subsets of predictors (see Table 4).

Non-adaptive LSTM surprisal (94.01% accuracy) and adaptive LSTM surprisal (94.06%) yielded the best classification accuracies when no other predictors were included. Over a baseline model comprised of every other feature except lexical repetition surprisal (see *base2* in Table 4), adaptive LSTM surprisal induced a small but significant increase of 0.03% in accuracy (p = 0.04 using McNemar's two-tailed test). When we included lexical repetition surprisal in the baseline model (see *base1* in Table 4), adaptive LSTM surprisal ceased to be a significant predictor. This suggests that, in the general case, the maximization of discourse predictability is driven by localized lexical priming captured by our trigram cache model. Apart from the content words, adaptive LSTM surprisal accounts for the re-occurrence of function words (e.g., case markers) which have been shown to modulate syntactic priming and drive parsing processes (Husain and Yadav, 2020).

To study prediction accuracy on non-canonical constructions, we restricted our analyses to IO- and DO-fronted items in the test partition (still training the classifier on the full training partition for each fold). In contrast to the DO-fronted subset, adaptive surprisal was a significant predictor

[6]The average dependency length difference for the DO-subset is 13.92 and for the IO-subset is 7.77 words.

Predictor	$\hat{\beta}$	$\hat{\sigma}$	t
intercept	**1.49**	0.008	171.18
trigram surp	**-0.28**	0.049	-5.84
dep length	**-0.05**	0.008	-6.22
pcfg surp	0.001	0.014	0.12
IS score	**0.04**	0.006	7.04
lex repetition surp	0.07	0.044	1.67
lstm surp	0.03	0.114	0.23
adaptive lstm surp	**-0.23**	0.113	-2.00

(a) Direct objects (DO; 1663 points) fronted cases

Predictor	$\hat{\beta}$	$\hat{\sigma}$	t
intercept	**1.51**	0.008	188.49
trigram surp	**-0.18**	0.039	-4.54
dep length	0.02	0.012	1.77
pcfg surp	**-0.13**	0.015	-8.34
IS score	-0.01	0.005	-1.87
lex repetition surp	0.03	0.036	0.92
lstm surp	**1.21**	0.154	7.87
adaptive lstm surp	**-1.50**	0.155	-9.67

(b) Indirect objects (IO; 1353 points) fronted cases

Table 3: Discourse adaptation regression model on DO/IO fronted cases (all significant predictors denoted by |t|>2)

Predictors	**Full Accuracy %**	**DO**	**IO**
a = IS score	51.84	53.88	50.92
b = dep length	62.31***	68.49***	58.91***
c = pcfg surp	86.86***	65.90	78.86***
d = lex repetition surp	90.07***	77.33***	85.07***
e = 3-gram surp	91.18***	78.95*	87.29**
f = lstm surp	**94.01*** **	79.55	87.28
g = adaptive lstm surp	94.06	**79.97**	**88.32*** **

Predictors	**Full Accuracy %**	**DO**	**IO**
Collective: with repetition effects			
base1 = a+b+c+d+e+f	**95.05**	80.99	89.06
base1 + g	95.06	81.06	**89.65***
Collective: without repetition effects			
base2 = a+b+c+e+f	95.06	81.24	89.65
base2 + g	**95.09***	81.42	89.80

Table 4: Prediction performances (Full data set (72833 points), Direct objects (DO; 1663 points) and indirect object (IO; 1353 points) fronted cases; each row refers to a distinct model; *** McNemar's two-tailed significance compared to model on previous row)

of IO-fronted syntactic choice, even in the presence of lexical repetition surprisal, as is evident from the significant increase of 0.6% in accuracy (p = 0.02 using McNemar's two-tailed test; see the rightmost IO column in Table 4). This result indicates that discourse predictability is effective in predicting IO-fronting in sentences that follow other IO-fronted sentences, suggesting the presence of syntactic priming effects. We consider adaptive LSTM LM surprisal (i.e., updating LM weights on successive sentences at test time) as an indicative of syntactic priming in this work but not the vanilla LSTM LM surprisal. We present a more nuanced discussion on this theme in Section 6.

Both our regression and classification results demonstrate that discourse adaptation is more effective in IO-fronted than DO-fronted constructions, mirroring the findings in Hindi sentence comprehension, where Vasishth (2004) showed that discourse context could compensate for the processing difficulty induced by indirect object fronting. The findings of our computational modelling reported in Table 4 are further validated by the agreement accuracy of our human evaluation study described in Section 3. Participants were more prone to prefer IO-fronted construction (80%) compared to DO-fronted construction (65%) as shown in Table 9 of Appendix G.

4.3 Qualitative Analysis: Success of Adaptive LSTM Surprisal

Further linguistic analyses in IO-fronted constructions revealed that LSTM adaptation also captured the priming of *given-given* items, potentially modeling the preferred ordering of multiple *given* items, a case not captured by IS score or lexical repetition surprisal. Reference sentence 1a is correctly predicted by the model containing adaptive LSTM surprisal and all other features (i.e., *base1+g* in Table 4) but a model without adaptive LSTM surprisal (i.e., base1) predicts the variant Example 1b. Appendix E Table 6 presents the exact scores of different predictors for the referent-variant pairs (1a and 1b). All predictors but LSTM and adaptive LSTM surprisal assign high score for the reference sentence with respect to its paired variant. Adaptive LSTM surprisal assigns a low per-word surprisal for the phrase *amar ujala* when it comes at

the first position in the reference sentence (1a) with respect to when it comes at the second position in the variant (1b), potentially modeling *givenness* as this word occurred in the previous sentence (Example 2 in Appendix E) as well. See Figure 3 in Appendix E for the information profile of the reference-variant pairs.

4.4 What causes priming?

In the priming literature, there is debate as to whether priming is driven by residual neural activation (short-lived effects) or by humans learning and updating their language expectations (long-lived effects). Bock and Griffin (2000) showed that syntactic priming in humans persisted even when prime and target sentences were separated by 10 intervening sentences, supporting the implicit learning (long-lived) hypothesis of syntactic priming. In order to test this effect on constituent ordering choice, we repeated our adaptation experiment by adapting to additional context sentences from the preceding discourse. Adaptive LSTM surprisal and lexical repetition surprisal were estimated by adapting the base LSTM LM and trigram LM, respectively, to five preceding context sentences, rather than the single context sentence we used for our other analyses. We found that for non-canonical IO/DO-fronted constructions, additional context sentences do not improve the adaptive LSTM LM's word order predictions, suggesting that priming may be driven by short-term residual activation (see Table 8 in the Appendix F).

5 Variance Inflation Factor

In this section, we evaluate our regression models for multicollinearity in terms of variance inflation factor (VIF) score. As Figure 2 in Appendix C denotes, the adaptive LSTM surprisal measure has a high correlation with all other surprisal predictors, which raises some suspicion that estimates of effects of the variables in our regression model might be unreliable. Table 10 in Appendix H presents the VIF scores for different regression models containing different set of predictors on full dataset, DO- and IO-fronted subsets. The outcomes indicate that all the surprisal predictors except PCFG surprisal in our original regression models have very high VIF scores (see Table 10a in Appendix H). Nevertheless, removing the more correlated measures, such as trigram surprisal and base LSTM surprisal does not alter any of our old results (see Table 11 for new regression results and Table 10b for corresponding VIF scores in Appendix H). All the surprisal measures including adaptive LSTM and lexical repetitions surprisal have negative regression coefficients suggesting that Hindi tends to optimize for discourse predictability.

6 Discussion

Our main findings suggest that in written Hindi, people choose word orders that maximize discourse predictability. The actual psychological mechanisms are conceivably lexical and structural priming. Our results indicate that lexical priming is most influential in canonical sentence contexts, but syntactic priming does influence ordering preferences in non-canonical contexts. Below, we discuss the implications of our findings in terms of the 4 factors affecting syntactic priming discussed in detail by Reitter et al. (2011): *inverse frequency interaction, decay, lexical boost, and cumulativity*. The IO-fronted construction is very rare (0.76% of our data) compared to DO-fronted non-canonical sentences (1% of our data) in the HUTB corpus of 13274 sentences. We find strong priming effects in IO-fronted constructions but weak priming in the case of DO-fronted constructions, providing evidence for an *inverse frequency interaction* (Scheepers, 2003; Jaeger and Snider, 2007).

Our finding that priming is not aided by long-term contexts indicates a *decay effect* in priming, which supports the residual activation (short-lived) hypothesis of priming in comprehension (Pickering and Branigan, 1998). Nevertheless, there has been evidence for implicit learning effects in comprehension as well (Luka and Barsalou, 2005; Wells et al., 2009). More recently, Ranjan et al. (2022b) using a similar setup as our current work argued for the existence of both the accounts *viz.,* residual activation and implicit learning, and demonstrated the role of dual mechanism priming effects (Tooley and Traxler, 2010) in Hindi word order.

Previous work suggests that lexical overlap between prime and target sentences enhances syntactic priming (Pickering and Branigan, 1998; Gries, 2005). The repeated lexical items become cues during sentence planning and bias the speaker to produce similar structures that those repeated lex-

ical items tend to occur in. Overall, we find that lexical repetition influences Hindi syntactic choice; however, syntactic priming is observed over and above lexical repetition in non-canonical constructions. It's interesting to note that comparable results have also been reported for English dialogue corpora (Healey et al., 2014; Green and Sun, 2021). We plan to conduct a systematic investigation on Hindi spoken data as a part of future work.

Finally, with regards to the *cumulativity* of priming, Jaeger and Snider (2007) showed in their corpus study of production of passives and *that*-insertion/omission that the effect of priming increases with the number of primes preceding it. Our work does not investigate this specifically, and more controlled experiments would be required.

The success of LSTM-based surprisal estimates over and above dependency length can also be interpreted in light of Futrell's (2019) point about the limitation of Surprisal Theory with respect to word order. Futrell modified Surprisal Theory by positing that the per-word processing difficulty is proportional to its surprisal given a *lossy memory representation* of the preceding context. Moreover, Futrell et al. (2020) proposed the Information Locality Hypothesis (ILH) which states that all pairs of words with high mutual information (not merely syntactically related words) tend to be located close to one another. The long window offered by LSTM surprisal thus models relationships between words at varying distances (over and above conventional trigram models). The success of these surprisal estimates for the task of reference sentence prediction provides some preliminary evidence for ILH in the case of word order.

Future work needs to tease apart priming effects of both vanilla LSTM and adaptive LSTM surprisal in the light of recent works. In this work, sentences are treated as independent while estimating their surprisal using vanilla LSTM LM, so there is no way vanilla LSTM can exhibit syntactic priming given the independent sentences. However, Misra et al. (2020) demonstrated that BERT exhibits "priming effect". The BERT LM was able to predict a word with greater probability when the context included a related word than an unrelated word. However, the effect decreased as the amount of information provided by the context increases. In other words, the related prime under high contextual constraint started acting as distractor—actively demoting the target word in the probability distribution; thus exhibiting "mispriming effect" (Kassner and Schütze, 2020). This could be due to stylistic avoidance of repeated structures/words in the adjacent sentences. Future work also needs to investigate whether word-order preferences can be jointly optimized using multiple factors (Gildea and Jaeger, 2015). In particular, the relationship between the drive to minimize surprisal (as found in this work) and the tendency to make information profiles uniform (Jaeger, 2010) needs to be explored more thoroughly in the light of recent findings (Meister et al., 2021).

Overall, our results demonstrate that Hindi word order preferences are influenced by discourse predictability maximization considerations. The actual mechanisms of discourse effects are plausibly lexical and syntactic priming.

7 Limitations

The 'levels' problem discussed in Levy (2018) which posits 2 levels of linguistic optimisation is germane while evaluating our work. Our results are restricted to the level of syntactic choices made by individual speakers or users of a given language over a lifetime (and not at the level of entire grammars and evolutionary timescales). Our experiments conducted on written text need to be performed on spoken data in order to make claims about priming in language production.

Acknowledgements

We thank the first author's dissertation committee members, Drs. Mausam and Samar Husain, as well as the Cornell C.Psyd members for their insightful comments and suggestions on this work. We thank Rupesh Pandey's logistical assistance in gathering the human judgment data for this work. We are also grateful for the thorough feedback provided by the anonymous reviewers of EMNLP 2022, ACL ARR 2021, and COMCO 2021. Finally, the last two authors also thank extramural funding from the Department of Science and Technology of India through the Cognitive Science Research Initiative (project no. DST/CSRI/2018/263).

References

Manabu Arai, Roger PG Van Gompel, and Christoph Scheepers. 2007. Priming ditransitive structures in comprehension. *Cognitive psychology*, 54(3):218–250.

Paul Baker, Andrew Hardie, Tony McEnery, Hamish Cunningham, and Robert Gaizauskas. 2002. Emille: a 67-million word corpus of indic languages: data collection, mark-up and harmonization. In *Proceedings of LREC 2002*, pages 819–827. Lancaster University.

Akshar Bharati, Rajeev Sangal, Vineet Chaitanya, Amba Kulkarni, Dipti Misra Sharma, and KV Ramakrishnamacharyulu. 2002. Anncorra: building tree-banks in indian languages. In *COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization*.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for Hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 186–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bock and Z. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning. *Journal of Experimental Psychology*, 2(120):177–192.

J.Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355 – 387.

Morten H. Christiansen and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509.

P.R. Clarkson and A. J. Robinson. 1997. Language model adaptation using mixtures and an exponentially decaying cache. In *In Proceedings of ICASSP-97*, pages 799–802.

Samvit Dammalapati, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2021. Effects of Duration, Locality, and Surprisal in Speech Disfluency Prediction in English Spontaneous Speech. In *Proceedings of the Society for Computation in Linguistics*, volume 4, page 10.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Vera Demberg, Asad B. Sayeed, Philip J. Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 356–367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alex B Fine and T Florian Jaeger. 2016. The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9):1362.

Alex B Fine, T Florian Jaeger, Thomas A Farmer, and Ting Qian. 2013. Rapid expectation adaptation during syntactic comprehension. *PloS one*, 8(10):e77661.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics.

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

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 199–206, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edward Gibson. 2000. Dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, brain: Papers from the First Mind Articulation Project Symposium*. MIT Press, Cambridge, MA.

Edward Gibson, Richard Futrell, Steven T. Piandadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Daniel Gildea and T. Florian Jaeger. 2015. Human languages order information efficiently. *CoRR*, abs/1510.02823.

Clarence Green and He Sun. 2021. Global estimates of syntactic alignment in adult and child utterances during interaction: Nlp estimates based on multiple corpora. *Language Sciences*, 85:101353.

Stefan Th. Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(4):365–399.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*. Cambridge University Press, New York.

John A. Hawkins. 2000. The relative order of prepositional phrases in english: Going beyond manner-place-time. *Language Variation and Change*, 11(03):231–266.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press.

John A. Hawkins. 2014. *Cross-Linguistic Variation and Efficiency*. Oxford University Press.

Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6):e98598.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Samar Husain and Himanshu Yadav. 2020. Target complexity modulates syntactic priming during comprehension. *Frontiers in Psychology*, 11:454.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage information density. *Cognitive Psychology*, 61(1):23–62.

T. Florian Jaeger and Elizabeth Norcliffe. 2009. The cross-linguistic study of sentence production: State of the art and a call for action. *Language and Linguistic Compass*, 3(4):866–887.

T. Florian Jaeger and Neal Snider. 2007. Implicit learning and syntactic persistence: Surprisal and cumulativity. *University of Rochester Working Papers in the Language Sciences*, 3:26–44.

T. Florian Jaeger and Harold Tily. 2011. Language processing complexity and communicative efficiency. *WIRE: Cognitive Science*, 2(3):323–335.

Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2018. Uniform Information Density Effects on Syntactic Choice in Hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48, Santa Fe, New-Mexico. Association for Computational Linguistics.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Y. Kachru. 2006. *Hindi*. London Oriental and African language library. John Benjamins Publishing Company.

Elsi Kaiser and John C Trueswell. 2004. The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94(2):113–147.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Roland Kuhn and Renato De Mori. 1990. A cache-based natural language model for speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 12(6):570–583.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126 – 1177.

Roger P Levy. 2018. Communicative efficiency, uniform information density, and the rational speech act theory.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171 – 193.

Barbara J Luka and Lawrence W Barsalou. 2005. Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, 52(3):436–459.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring bert's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin J. Pickering and Holly P. Branigan. 1998. The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language*, 39(4):633–651.

Ting Qian and T. Florian Jaeger. 2012. Cue effectiveness in communicatively efficient discourse production. *Cognitive Science*, 36(7):1312–1336.

Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler. 2016. Investigating locality effects and surprisal in written english syntactic choice phenomena. *Cognition*, 155:204–232.

Rajakrishnan Rajkumar and Michael White. 2014. Better surface realization through psycholinguistics. *Language and Linguistics Compass*, 8(10):428–448. ISSN: 1749-818X.

Sidharth Ranjan, Sumeet Agarwal, and Rajakrishnan Rajkumar. 2019. Surprisal and Interference Effects of Case Markers in Hindi Word Order. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2022a. Locality and expectation effects in hindi preverbal constituent ordering. *Cognition*, 223:104959.

Sidharth Ranjan, Marten van Schijndel, Sumeet Agarwal, and Rajakrishnan Rajkumar. 2022b. Dual mechanism priming effects in hindi word order. *arXiv preprint arXiv:2210.13938*.

David Reitter, Frank Keller, and Johanna D. Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive Science*, 35(4):587–637.

Rajeev Sangal, Vineet Chaitanya, and Akshar Bharati. 1995. *Natural language processing: a Paninian perspective*. PHI Learning Pvt. Ltd.

C Scheepers. 2003. Syntactic priming of relative clause attachments: persistence of structural configuration in sentence production. *Cognition*, 89:179–205.

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Andreas Stolcke. 2002. SRILM — An extensible language modeling toolkit. In *Proc. ICSLP-02*.

David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

Kristen M Tooley and Matthew J Traxler. 2010. Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, 4(10):925–937.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710.

S. Vasishth. 2004. Discourse context and word order preferences in Hindi. *Yearbook of South Asian Languages*, pages 113–127.

Justine B Wells, Morten H Christiansen, David S Race, Daniel J Acheson, and Maryellen C MacDonald. 2009. Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, 58(2):250–271.

Marta Wierzba and Gisbert Fanselow. 2020. Factors influencing the acceptability of object fronting in german. *The Journal of Comparative Germanic Linguistics*, 23(1):77–124.

Himanshu Yadav, Ashwini Vaidya, and Samar Husain. 2017. Keeping it simple: Generating phrase structure trees from a Hindi dependency treebank. In *TLT*.

Appendix

A Variant Generation

(2) **Context sentence**

amar ujala-ki bhumika nispaksh rehti
Amar Ujala-GEN role unbiased remain
hai
be.PRS.SG

Amar Ujala's role remains unbiased.

(3) a. amar ujala-ko **yah** *sukravar*-ko
Amar Ujala-ACC **it** *friday*-on
daak-se prapt hua
post-INST receive be.PST.SG
[Given-Given = 0] **(Reference)**

Amar Ujala received **it** by post on *Friday*.

b. **yah** amar ujala-ko *sukravar*-ko daak-se prapt hua [Given-Given = 0] **(Variant 1)**

c. *sukravar*-ko **yah** amar ujala-ko daak-se prapt hua [New-Given = -1] **(Variant 2)**

This work uses sentences from the Hindi-Urdu Treebank (HUTB) corpus of dependency trees (Bhatt et al., 2009) containing well-defined subject and object constituents. Figure 1 displays the dependency tree (and a glossary of relation labels) for reference sentence 3a. The grammatical variants were created using an algorithm that took as input the dependency tree corresponding to each HUTB reference sentence. The reordering algorithm permuted the preverbal[7] dependents of the root verb and linearized the resulting tree to obtain variant sentences. For example, corresponding to the reference sentence 3a and its root verb "hai" (see figure 1a), the preverbal constituents with parents as "ujala", "yah", "suravar", "daak", and "prapt" were permuted to generate the artificial variants (3b and 3c). The ungrammatical variants were automatically filtered out using dependency relation sequences (denoting grammar rules) attested in the gold standard corpus of HUTB trees. In the dependency tree 1a, "k4-k1", "k7t-k1", "k3-k7t", and "pof-k3" are dependency relation sequences. In cases where the total number of variants exceeded 100 (a random cutoff),[8] we chose 99 non-reference variants randomly along with the reference sentence.

[7]Hindi is not a strictly verb-final language, but the majority of the constituents in the HUTB corpus are preverbal. Our corpus analysis of 13274 sentences present in HUTB suggests 20,750 pairs of preverbal constituents and 2599 pairs of postverbal constituents. Therefore, our variant generation (via reordering of constituents) and subsequent experiments focus on word-order variation in the preverbal domain, considering the preverbal domain to be the locus of word-order variation. Only preverbal constituents are permuted to generate grammatical variants and leave the postverbal constituents in the reference-variants sentences as it is.

B Information Status Annotation

The subject and object constituents in a sentence were assigned a *Given* tag if any content word within them was mentioned in the preceding sentence or if the head of the phrase was a pronoun. All other phrases were tagged as *New*. The sentence example 3 illustrates the proposed annotation scheme.

- Example 3a follows *Given-Given* ordering — The object "Amar Ujala" in the sentence is mentioned in the preceding context sentence 2, it would be annotated as *Given*. In contrast, the subject "yah" is a pronoun so it would also be tagged as *Given* following the annotation scheme.

- Example 3c follows *New-Given* ordering — The object "sukravar" in the sentence should be tagged as *New* as it is not mentioned in the preceding context sentence 2. In contrast, the subsequent pronoun "yah", which acts as the subject of the sentence, should be tagged as *Given* following the annotation scheme.

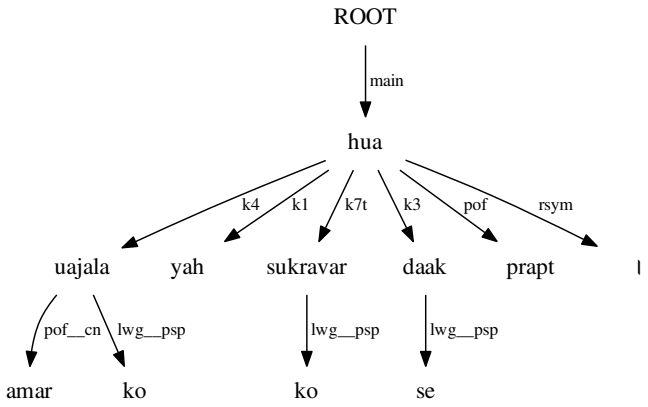
C Correlation Plot

The Pearson's correlation coefficients between different predictors are displayed in Figure 2. The adaptive LSTM surprisal has a high correlation with all other surprisal features and a low correlation with dependency length and information status score.

D Joachims Transformation

This technique converts a binary classification problem into a pair-wise ranking task involving the

[8]Higher and lower cutoffs do not affect our results.



(a) Dependency tree

Label	Dependency relation
Invariant syntactic relations	
k1	subject/agent
k2	object/patient
k3	instrument
k4	object/recipient
k7t	location in time
Complex predicate relation	
pof	parts of conjunct verb
pof_cn	parts of compound noun
Local word group (lwg)	
lwg_psp	postposition
lwg_vaux	auxilliary verb
Symbols	
rsym	symbol relation

(b) Dependency relations

Figure 1: Example HUTB dependency tree and relation labels

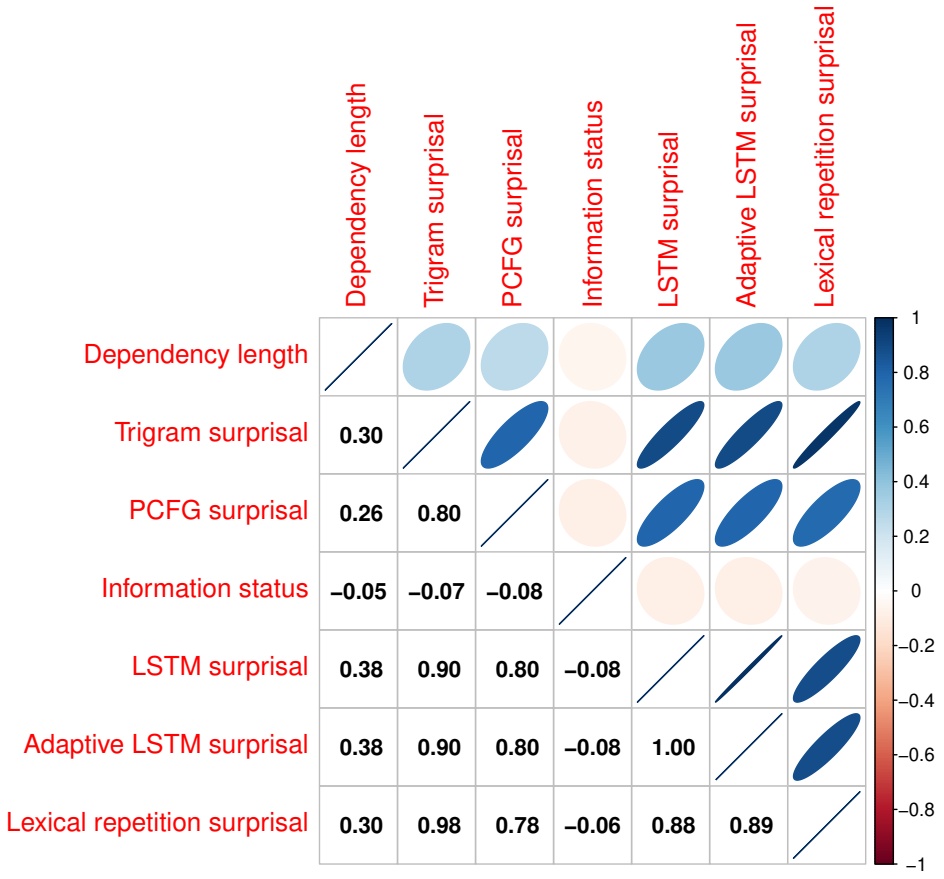


Figure 2: Pearson's coefficient of correlation between different pairs of predictors

feature vectors of a reference sentence and each of its variants. Table 5 displays the Joachims's transformation. The delta (δ) refers to the difference between the feature vectors of the reference sentence and its paired variant. The overall goal is to model two-alternative choices for each reference sentence such that the speaker generates the reference sentence after rejecting a potential grammatical variant. Moreover, the reference sentence that appeared originally in the corpus must have been present due to its properties (*viz.,* dependency length, discourse context, accessibility, or surprisal), and variant sentences are less likely to be produced due to the same reasons.

E Information Profile for IO-fronted Example

The LSTM LM, when adapted to the previous sentence (2) in the discourse, assigns a lower surprisal score to the *given* item when it occurs in the first position ("amar ujala" in sentence 3a) than when it appears in the second position ("amar ujala" in sentence 3b) in the subsequent sentence.

F Contextual Adaptation on One Vs. Multiple Sentences for DO/IO Constructions

We investigated if adapting the LSTM LM to the preceding five contextual sentences instead of one contextual sentence will help predict word-ordering patterns better for IO/DO constructions. Table 7 showcases perplexity dip on test sentences during 1 vs. 5 contextual sentence adaptation. Table 8 highlights the classification accuracy of different models containing a combination of features. Our results indicate that adding *Prev5-adaptive* LSTM surprisal in the machine learning model above and beyond every other feature, including *Prev1-adaptive* surprisal (*i.e.,* Baseline) does not significantly boost prediction accuracy for both IO- and DO-fronted subset. A similar finding was observed when Prev5-lexical-repetition surprisal (*Prev5-Lex-Rept LM*: base trigram LM interpolated with unigram cache LM containing 5 preceding sentences history) was included in the classifier model above and beyond every other feature (*i.e.,* Baseline) including *Prev1-lexical-repetition* surprisal (*Prev1-Lex-Rept LM*: base trigram LM interpolated with unigram cache LM containing only 1 preceding sentence history).

G Human Evaluation

To determine whether the permuted word order (variant) is dispreferred to the original word order (reference), we conducted a targeted human evaluation via a forced-choice task and collected sentence judgments from 12 Hindi native speakers for 167 randomly selected reference-variant pairs in our data set. Participants were first shown the preceding sentence and then asked to judge the subsequent most likely sentence as the best choice between the reference-variant pair. Each sentence was assigned a human label of "1" if more than 50% participants voted it as best choice else human label of "0". The stimuli belonged to two different constructions, viz., the reference sentence (Ref) has canonical ordering whereas, the variant (Var) has non-canonical ordering (DO-fronted or IO-fronted) and vice versa.

Table 9 presents the results. On the entire dataset containing 167 reference-variant pairs, 89.92% (agreement accuracy) of the reference sentences originally appearing in the HUTB corpus were also preferred by native speakers compared to the artificially generated grammatical variants expressing similar meanings. Moreover, as initially hypothesized, the Hindi participants were more prone to prefer IO-fronted construction (80%) compared to DO-fronted construction (65%) as captured by the agreement accuracy validating the findings reported in Table 4. Overall, the full model containing all the features, including adaptive LSTM surprisal, predicted human preferences (76.65%) much better than corpus choice labels (74.85%).

H Variance Inflation Factor Analysis

Table 10a displays the VIF scores for each predictor in the different regression models. The VIF scores for the regression models without the correlated features, such as trigram surprisal and vanilla LSTM surprsial, are documented in Table 10b. And Table 11 reports the results of the regression experiment when the model did not contain these highly correlated features.

Condition	Label	Dependency length	n-gram surprisal	pcfg surprisal
Reference	1	18	24.69	61.13
Variant$_1$	0	20	23.80	60.67
Variant$_2$	0	18	23.02	60.02

(a) Original feature values

Condition	Label	δ Dependency length	δ n-gram surprisal	δ pcfg surprisal
Variant$_1$ -Reference	0	2	-0.89	-0.46
Reference -Variant$_2$	1	0	1.67	1.11

(b) Transformed feature values

Table 5: Joachims transformation

Type		3g surp	Deplen	PCFG surp	IS score	LSTM surp	Adaptive LSTM surp	Lex rept surp
Example 3a	Reference	24.69	18	61.13	0	**91.80**	**89.52**	23.80
Example 3b	Variant	23.80	20	60.67	0	93.78	93.17	22.19

Table 6: Predictor scores for reference-variant pairs (3a, 3b)

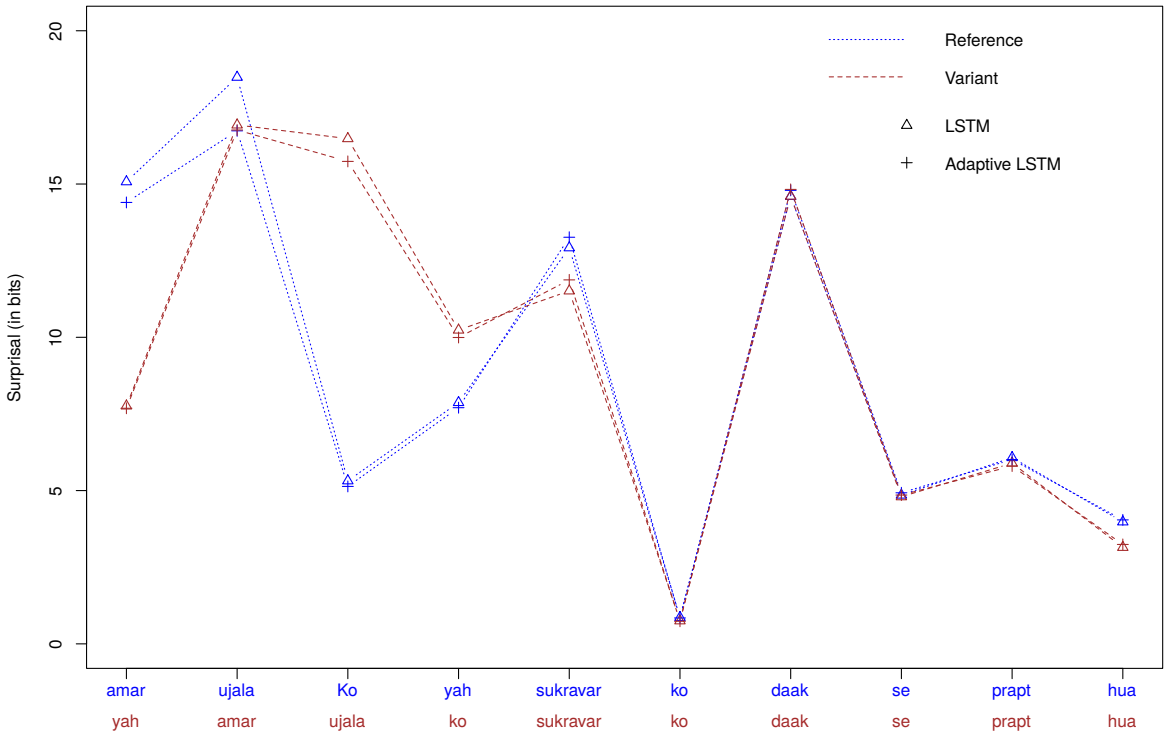


Figure 3: Information profile for the reference-variant pair 3a and 3b

Non-canonical HUTB Sentences	**Frequency Count (%)**	**Baseline Perplexity**	**Adapted Perplexity (Prev1)**	**Perplexity Dip (Prev1)**	**Adapted Perplexity (Prev5)**	**Perplexity Dip (Prev5)**
DO	133 (1%)	183.92	103.40	-80.52	77.33	-106.59
IO	101 (0.76%)	138.78	88.26	-50.52	68.45	-70.33

Table 7: Effect of adaptation on discourse sentences (Prev1: Preceding one sentence in discourse, Prev5: Preceding five sentences in discourse)

Type	Baseline	+ Prev5 Adaptive LSTM Surp	+ Prev5 Lex-Rept Surp
DO	81.06	81.12	80.94
IO	89.65	89.73	89.51

Table 8: Prediction performance (Direct objects (DO: 1663 points), Indirect Objects (IO: 1353 points)); Baseline denotes *base1+g* shown in Table 4; bold denotes McNemar's two-tailed significance compared to a baseline model in the same row

Construction Type (item count)	Agreement (%) human:corpus	Model (%) corpus labels	Model (%) human labels
Ref: Canonical | Var: DO-fronted (20)	95	90	85
Ref: DO-fronted | Var: Canonical (20)	**65**	25	50
Ref: Canonical | Var: IO-fronted (20)	100	85	85
Ref: IO-fronted | Var: Canonical (20)	**80**	65	65
Ref: Non-canonical | Var: Canonical (80)	85.00	66.25	71.25
Ref: Canonical | Var: Canonical (87)	94.25	82.76	81.61
Total (167)	89.92	74.85	76.65

Table 9: Targeted human evaluation — **Agreement human/corpus**: Percentages of times human judgment matches with corpus reference choice; **Model corpus**: Percentages of corpus choice correctly predicted by the classifier containing all the predictors (*base1 + g* as per Table 4); **Model human:** Percentages of human label correctly predicted by the classifier containing all the predictors (*base1 + g* as per Table 4)

Predictors	Full (72833)	DO (1663)	IO (1353)
trigram surp	**27.61**	18.87	18.87
dependency length	1.18	1.46	1.46
pcfg surp	3.08	1.86	1.86
IS score	1.01	1.03	1.03
lex-rept surp	**24**	16.47	16.47
lstm surp	**241.62**	109.04	109.04
adaptive lstm surp	**244.37**	106.97	106.97
PERFORMANCE			
Residual standard err	0.271	0.294	0.354
Multiple R-squared	0.707	0.657	0.502
Adjusted R-squared	0.707	0.655	0.500

(a) All predictors

Predictors	Full (72833)	DO (1663)	IO (1353)
dependency len	1.18	1.42	1.18
pcfg surp	2.95	1.73	2.08
IS score	1.01	1.01	1.06
lex-rept surp	4.99	4.45	4.26
adaptive lstm surp	5.77	4.39	4.29
PERFORMANCE			
Residual std. err	0.271	0.357	0.304
Multiple R-sqrd	0.706	0.492	0.633
Adjusted R-sqrd	0.706	0.490	0.631

(b) Predictors except trigram and lstm surprisal measures

Table 10: Variance inflation factor analysis on different regression models containing: (a) all predictors b) all predictors but correlated features; Each column denotes individual models on a given dataset with a different set of predictors; VIF larger than 5 or 10 indicates that the model has problems estimating the coefficient of variables

Predictor	$\hat{\beta}$	$\hat{\sigma}$	t
Intercept	1.5	0.001	1493.53
dependency len	0.02	0.0011	15.84
pcfg surp	-0.08	0.0017	-43.49
IS score	0.01	0.001	11.92
lex-rept surp	-0.09	0.0022	-39.97
adaptive lstm surp	-0.28	0.0024	-116.06

Table 11: Regression model on full data set after removing the correlated features ($N = 72833$; all significant predictors denoted by |t|>2)