

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Does Dependency Locality Predict Non-canonical Word Order in Hindi?

Permalink

<https://escholarship.org/uc/item/71k6d904>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Ranjan, Sidharth
van Schijndel, Marten

Publication Date

2024

Peer reviewed

Does Dependency Locality Predict Non-canonical Word Order in Hindi?

Sidharth Ranjan (sidharth.ranjan@ling.uni-stuttgart.de)

University of Stuttgart, 70174 Stuttgart, Germany

Marten van Schijndel (mv443@cornell.edu)

Cornell University, 14853 Ithaca, United States

Abstract

Previous work has shown that isolated non-canonical sentences with Object-before-Subject (OSV) order are initially harder to process than their canonical counterparts with Subject-before-Object (SOV) order. Although this difficulty diminishes with appropriate discourse context, the underlying cognitive factors responsible for alleviating processing challenges in OSV sentences remain a question. In this work, we test the hypothesis that dependency length minimization is a significant predictor of non-canonical (OSV) syntactic choices, especially when controlling for information status such as givenness and surprisal measures. We extract sentences from the Hindi-Urdu Treebank corpus (HUTB) that contain clearly-defined subjects and objects, systematically permute the preverbal constituents of those sentences, and deploy a classifier to distinguish between original corpus sentences and artificially generated alternatives. The classifier leverages various discourse-based and cognitive features, including dependency length, surprisal, and information status, to inform its predictions. Our results suggest that, although there exists a preference for minimizing dependency length in non-canonical corpus sentences amidst the generated variants, this factor does not significantly contribute in identifying corpus sentences above and beyond surprisal and givenness measures. Notably, discourse predictability emerges as the primary determinant of constituent-order preferences. These findings are further supported by human evaluations involving 44 native Hindi speakers. Overall, this work sheds light on the role of expectation adaptation in word-ordering decisions. We conclude by situating our results within the theories of discourse production and information locality.

Keywords: Word order; Discourse; Locality; Production

Introduction

The SUBJECT (S), OBJECT (O), and VERB (V) collectively constitute the building blocks or constituents of a sentence, operating together in a structured manner to convey the intended meaning. However, the linguistic properties of each language result in a spectrum of legitimate constituent orders within the sentence which all potentially express the same idea (Chomsky, 1965; Ferreira, 1996), some of these are more common than others. The most frequent constituent order is often called *canonical* and is often considered the most natural and basic syntactic structure in a language. It is typically considered neutral and unmarked. In contrast, the alternative *non-canonical* orders are rarer and are employed under specific information structural and discourse considerations (Comrie, 1981; Gambhir, 1981; Bock, 1982; Mohanan & Mohanan, 1994). Comprehenders often have greater difficulty understanding these less frequent constructions than those that are commonly produced (Hale, 2001; Levy, 2008;

Hawkins, 2014). However, the same rare non-canonical orders, when used within an appropriate discourse or situational context, can be as preferable and easy to process as sentences with canonical orders (Kaiser & Trueswell, 2004; Hörnig, Oberauer, & Weidenfeld, 2005; Kristensen, Engberg-Pedersen, & Poulsen, 2014; Bader & Meng, 2023).

While Hindi, a flexible word order language (Kachru, 1982), usually employs SOV order, Vasishth, Shaher, and Srinivasan (2012) found that context rendered certain non-canonical word orders easier to comprehend than their canonical counterparts, suggesting that syntactic complexity can be mitigated by context. In spite of this, Vasishth (2004) has shown that some non-canonical OSV sentences such as 2a remain harder to comprehend than 2b despite the presence of an appropriate discourse context (1). This difficulty seems to arise due to a considerably larger distance between the fronted object and the verb, thereby increasing the total dependency length within sentence (Gibson, 2000).

(1) Discourse context sentence¹

air india aur indian vimaan sevaayon-ko notice de
Air India and Indian airlines services-ACC notice give
diya gya hai
give.PRS.PFV.SG

Air India and Indian Airlines services have been notified.

(2) Target sentence

- a. [yah notice]DO [kuwait-ke vimaan pradhikaran
this notice kuwait-GEN aviation authority
mantralaye-ke director general-ne]S jaari [kiya]V
ministry-GEN general-ERG issue do
hai (Reference; OSV; Given-Given)
be.PRS.PFV.SG

This notice has been issued by the Director-General of the Ministry of Aviation Ministry of Kuwait.

- b. [kuwait-ke vimaan pradhikaran mantralaye-ke direc-
tor general-ne]S [yah notice]DO jaari [kiya]V hai
(Variant; SOV; Given-Given)

In this work, we study the cognitive factors that influence a speaker's choice to produce a non-canonical order when an equally viable canonical order exists in Hindi. While naturally occurring non-canonical sentences do minimize dependency length as compared with the canonical alternatives, we find that givenness and surprisal minimization are stronger predictors of this behavior.

¹Major constituents are annotated with brackets, repeated content words are underlined in both context and target sentences.

Hypothesis

In a recent study, Ranjan, Rajkumar, and Agarwal (2022) demonstrated that although dependency length minimization is generally a weak predictor of Hindi word order preferences, it significantly predicts non-canonical (OSV) orders, even in presence of surprisal estimates from trigram and incremental dependency parsing models. They argued that surprisal measures may perform poorly for non-canonical orders since the models for estimating surprisal are themselves biased towards more frequent structures, while dependency length minimization proved effective because fronted-objects were consistently longer than the subjects, resulting in sequences favoring long-before-short patterns in the corpus.

Following their work, we set out to test the hypothesis that dependency length minimization is a significant predictor of non-canonical (OSV) production choices, especially when controlling for information status such as givenness and surprisal. To do this we use dependency length, givenness, and a variety of surprisal measures to predict the likelihood of non-canonical vs canonical constructions in naturally occurring written sentences from the Hindi-Urdu Treebank corpus (HUTB; Bhatt et al., 2009). If a factor influences production decisions in humans, it should be able to correctly anticipate where different constituent orderings will occur in the corpus.

Measures of Processing Difficulty

Dependency Locality Theory

Dependency Locality Theory (DLT; Gibson, 1998, 2000) has proven effective both in predicting comprehension difficulty and in modeling production durations (Scontras, Badecker, Shank, Lim, & Fedorenko, 2014; Dammalapati, Rajkumar, Ranjan, & Agarwal, 2021).

DLT posits two types of processing costs: INTEGRATION COST and STORAGE COST. In general, shorter dependencies have been found to dominate across natural languages (Liu, 2008; Futrell, Levy, & Gibson, 2020; Ranjan & von der Malsburg, 2023, 2024). The placement of *long-before-short* constituents in SOV languages and *short-before-long* constituents in SVO languages is preferred as they minimize total dependency length within the sentence (Hawkins, 1994, 2004). In this study, we define dependency length as the count of intervening words between head and dependent units within a dependency graph, and calculate the sentence-level dependency length by summing each word’s dependency length within the sentence (Temperley, 2007).

Surprisal

Claude Shannon’s (1948) work on Information Theory defined the amount of new information contained in an observation, based on the (un)predictability of that observation before it occurred. Hale (2001) showed that this measure, called *surprisal*, can predict language processing behavior:

$$S_{k+1} = -\log P(w_{k+1}|w_{1..k}, C) \quad (1)$$

The *surprisal* of a word (w_{k+1}) is defined in terms of the conditional probability of that word, w given its intra-sentential context ($w_{1..k}$) and inter-sentential context (C ; Eq. 1). These probabilities can be estimated using either a *sequence of words* or a *syntactic tree* (Levy, 2008). Many studies have now shown that words with high surprisal are difficult to process and comprehend (Jurafsky, 1996; Hale, 2006; Demberg & Keller, 2008; Smith & Levy, 2013). In this work, we estimate per-word surprisal using a number of methods described below and then compute the sentence-level surprisal for each method by summing the per-word scores within each sentence.

1. ***n*-gram surprisal:** To capture linear local predictability, we used a traditional trigram language model to estimate the surprisal of the target word conditioned only on the two preceding context words. The trigram model we used was trained on 1 million sentences from the EMILLE corpus (Baker, Hardie, McEnergy, Cunningham, & Gaizauskas, 2002) with Good-Turing discounting using the SRILM toolkit (Stolcke, 2002).
2. **PCFG surprisal:** To capture structural predictability, we computed the conditional probability of the words given the incremental available syntactic structures. This metric was computed for each word in the sentence using the Modelblocks incremental constituency parser (van Schijndel, Exley, & Schuler, 2013).² Because there are no constituency treebanks for Hindi, we used the (Yadav, Vaidya, & Husain, 2017) method to generate 12,000 constituency trees from the dependencies in the HUTB corpus. Since we wanted to use these trees both to train our parser and to analyze in our study, we used a 5-fold cross-validation (CV) method whereby each fifth of our data was annotated with PCFG surprisal after training our parser on the other four-fifths of our data.³
3. **Adaptive LSTM surprisal:** To capture long-range predictability effects, we used the Neural Complexity Toolkit (van Schijndel & Linzen, 2018).⁴ This toolkit implements online fine-tuning of a neural language model, where an LSTM language model is updated on each context sentence to better model in influence of discourse context (Example 1), and then surprisal of the words in the target sentences (Example 2) are estimated using the revised language model weights. This approach generates discourse-sensitive surprisal estimates that considers not only intra-sentential unbounded context but also inter-sentential information. Using this method, van Schijndel and Linzen (2018) found that adaptive surprisal was significantly better at predicting human reading times than non-adaptive surprisal.

²<https://github.com/modelblocks>

³F1-score: 90.82% (train) and 84.95% (test) in 5-fold CV setting

⁴<https://github.com/vansky/neural-complexity>

| Dataset | Ordering type | Reference (%) | Variant (%) |
|--------------|----------------------|---------------|----------------|
| O-fronted | Ref: OSV; Var: SOV | 233 (11.67%) | 2834 (3.89%) |
| DO-fronted | Ref: DOSV; Var: SDOV | 133 (6.67%) | 1663 (2.28%) |
| IO-fronted | Ref: IOSV; Var: SIOV | 101 (5.06%) | 1353 (1.86%) |
| Canonical | Ref: SOV; Var: OSV | 1763 (88.33%) | 30851 (42.36%) |
| Full dataset | | 1996 | 72833 |

Table 1: Distribution of ordering types in Reference (Ref) and Variant (Var) sentences in our dataset; S = subject; O = object; V = verb; DO = direct object; IO = indirect object

Information Status

Languages are known to follow a principle of GIVEN-BEFORE-NEW ordering, by re-referring to information previously mentioned in the discourse *before* introducing the new information in the sentence (Clark & Haviland, 1977).

We annotate each sentence in our dataset with Information Status (IS). For this, we examine subject and object phrases in the target sentences (2), checking if any content words within these phrases are mentioned in the preceding context sentence (1). If so, these phrases are tagged as GIVEN, otherwise they are tagged as NEW. Additionally, we label phrases as GIVEN if the head of these constituents is a pronoun. Under this scheme, both reference (2a) and variant (2b) sentences display a GIVEN-GIVEN ordering as both the subject and the object contain content words (underlined) that are present in the context sentence (1). Numerical IS scores were assigned as follows: a) Given-New order = +1, b) New-Given order = -1, c) Given-Given and New-New orders = 0.

Data and Methods

In our study, we used sentences from the HUTB corpus of written newswire sentences (Bhatt et al., 2009). We only analyzed those trees that adhered to the following criteria: a) the trees contain both well-defined subjects and objects, b) the trees are projective, c) the sentences are declarative, and d) the root node for each tree is a finite verb with at least two preverbal dependents.

We generated counterfactual variants for each reference sentence in our dataset. This involved randomly permuting the preverbal constituents within the dependency tree, specifically those directly dependent on the root verb (see Ex. 2). The preverbal domain serves as the primary locus of constituent order variation in SOV languages, with fewer post-verbal constituents (Ranjan, Rajkumar, & Agarwal, 2022). We limited the variant generation to a maximum of 99 variants per corpus reference sentence as an arbitrary cutoff for computational tractability, though our findings remained consistent regardless of the chosen cutoff. Ungrammatical variants were automatically filtered out using grammar rules derived from the corpus dependency trees in the treebank (Rajkumar & White, 2014).

Our variant generation procedure resulted in a total of 72833 grammatical variants for 1996 reference sentences in the corpus. We then categorized reference and variants by their word orders into canonical (SOV) and non-canonical (OSV) types, as illustrated in Table 1. The occurrence dis-

tribution revealed that non-canonical reference sentences are indeed less frequent compared to canonical orders. Following this, we set up a classification task to investigate if the features described in the preceding section are effective in identifying the corpus reference sentences amidst competing counterfactual variants. However, the presence of a greater number of variants (72833) per reference sentence (1996) introduces a substantial class imbalance issue (reference / variant) for our binary classification task. The ranking method in the next section addresses this class imbalance problem. We then use the transformed dataset with balanced class labels to report our regression and classification analyses.

We also conducted a targeted human evaluation via a two-alternative forced choice (2AFC) judgment task involving 44 Hindi native speakers. We collected human judgments over 164 randomly chosen reference-variant pairs from our full dataset. Participants first viewed the context sentence (Ex. 1) and then selected the preferred continuation between the reference (Ex. 2a) and the variant (Ex. 2b). Results showed that, 85.63% of the HUTB reference sentences (out of 164 pairs) were preferred by native speakers over the plausible variants that we generated. As a result, we feel justified in considering the reference sentences in the HUTB corpus as gold standards of human ordering preferences in our analyses.

Ranking Model

Our original dataset contained significantly more variants (72833) than reference sentences (1996) and this imbalance posed a problem for training an unbiased classifier. To rectify this imbalance, we applied a data transformation technique originally proposed for ranking web pages by Joachims (2002). Previous work on syntactic choice phenomena has successfully employed this method, which converts a binary classification task with heavily imbalanced labels into a pairwise ranking task (Rajkumar, van Schijndel, White, & Schuler, 2016). This conversion associates feature vectors from the reference sentences with those of their variants. The feature vectors in this case consist of the aforementioned features which we computed for both reference and variant sentences. Next, we trained a logistic regression model on the difference between feature vectors of reference and variant pairs, as illustrated in the equations below:

$$\mathbf{w} \cdot \phi(\text{Reference}) > \mathbf{w} \cdot \phi(\text{Variant}) \quad (2)$$

$$\mathbf{w} \cdot (\phi(\text{Reference}) - \phi(\text{Variant})) > 0 \quad (3)$$

Eq. 2 depicts a standard classification model determining whether the reference sentence outperforms one of its variants. This decision relies on comparing the dot product of the feature vector linked to the reference sentence and the learned feature weights \mathbf{w} with the corresponding dot product of the variant sentence. This relationship is alternatively expressed in Eq. 3, where the feature values of the first pair member are subtracted from their counterparts in the second member (Joachims, 2002). The model’s decision for a specific reference-variant pair can then be made by evaluating the

sign of the dot product between the learned feature weights and the feature vector difference (Eq. 3).

We generated ordered pairs comprising feature vectors for both reference (REF) and variant (VAR) sentences, maintaining a balance in the counts of each order type (REF-VAR, VAR-REF). Pairs alternating between ‘REF-VAR’ were labeled as ‘1,’ whereas pairs in the ‘VAR-REF’ sequence were labeled as ‘0.’ This coding strategy ensures a balanced dataset, with a near equal number of labels for each type.⁵ The transformed features were then fed into a logistic regression model (using the `glm` function in R) to test our hypothesis:

$$choice \sim \begin{cases} \delta_0 \text{ adaptive lstm surprisal} + \\ \delta_1 \text{ trigram surprisal} + \delta_2 \text{ pcfg surprisal} + \\ \delta_3 \text{ IS score} + \\ \delta_4 \text{ dependency length} \end{cases} \quad (4)$$

Here, *choice* is a binary dependent variable (1 denotes a preference for the reference sentence, and 0 denotes a preference for the variant sentence). The deltas (δ_x) refer to the difference between each feature of the reference sentence and the paired variant. All the independent variables were normalised to z-scores.

To evaluate the classification performance, we examined the output of our model trained on the entire dataset (72833 data points) with transformed feature values as predictors using 10-fold cross-validation method i.e., models trained on 9 folds of the dataset were used for prediction in the remaining fold. However, for estimating regression coefficients in the model, we use the test data with transformed feature values as predictors for a given construction under study.

Results

In this section, we investigate whether dependency length minimization is a significant predictor of OSV syntactic choices, even in presence of surprisal and givenness measures. As discussed previously, corpus orderings were human-preferred compared with plausible alternatives. We assume that this is due to the psycholinguistic properties of each sentence. We first used regression to identify the strength with which each factor predicts the corpus orderings compared with possible but unattested alternatives. Then we examined the percentage of correctly predicted reference sentences over paired alternative variants using our features. Finally, we collected human preference data to determine how closely these features cause our automatic classifier to mimic human behavior.

Regression Analysis

In this section, we test our main hypothesis that dependency length predicts non-canonical OSV orders (i.e. DO-fronted and IO-fronted corpus sentences) while controlling for surprisal measures and IS score. Based on the prior literature, we expect that surprisal should be lower in the reference than in the variant and that givenness should be adhered to more

⁵The number of labels is equal when the total number of variants is even and they differ by one when the variant total is odd

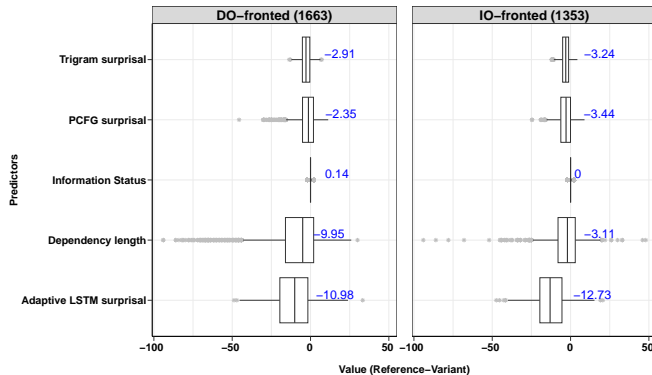


Figure 1: Summary of dataset denoting difference between predictor values of Reference and their corresponding variant sentence for DO-fronted and IO-fronted constructions; Mean difference value for each predictor is annotated.

in the reference sentence than the variant.⁶ If dependency length minimization plays a role in these ordering decisions, we would expect dependency length to be lower in the reference sentences compared with the variants. Our results (Figure 1) suggest that surprisal and dependency length predictions generally hold across both types of non-canonical constructions. However, the givenness (information status) prediction holds true only for DO-fronted sentences.

For the DO-fronted subset of our corpus (Table 2), reference sentences consistently minimize dependency length in addition to trigram and adaptive LSTM surprisal as shown by their negative regression coefficients. However, the positive regression coefficient for PCFG surprisal suggests greater syntactic complexity associated with DO-fronted constructions. The positive regression coefficient for IS score implies that DO-fronted corpus sentences exhibit GIVEN-NEW ordering as compared to the paired alternative variant. Adding dependency length to a baseline model containing all other predictors significantly improved the fit of our model in the DO-fronted subset ($\chi^2 = 51.61$; $p < 0.001$). In contrast, for IO-fronted constructions (Table 3), we observe that reference sentences consistently minimize all surprisal measures as evinced by their negative regression coefficients. Dependency length and IS score were not significantly predictive of IO-fronted orderings.

Our regression results are largely in line with previous findings in the literature, which have found differential processing effects of dependency length for DO- and IO-fronted constructions in Hindi (Vasisht, 2004; Ranjan, Rajkumar, & Agarwal, 2022; Ranjan, van Schijndel, Agarwal, & Rajkumar, 2022). Ranjan and colleagues, in their analysis of HUTB sentences found a preference for *long-before-short* constituent placement in DO-fronted sentences in line with dependency length minimization (DLM). In contrast, they

⁶Based on the coding of given-new as +1 and new-given as -1, we expect the reference-variant difference [(+1) - (-1) = 2] for IS score to be a positive value if givenness is truly adhered to.

| Predictors | Estimate | Std. Error | Z-value | VIF |
|-------------------------|--------------|------------|----------|------|
| Intercept | 0.02 | 0.074 | 0.25 | – |
| Dependency length | -0.58 | 0.091 | -6.35*** | 1.02 |
| IS score | 0.26 | 0.055 | 4.68*** | 1.03 |
| PCFG surprisal | 0.52 | 0.135 | 3.88*** | 1.42 |
| Trigram surprisal | -2.08 | 0.245 | -8.48*** | 2.12 |
| Adaptive LSTM surprisal | -1.65 | 0.240 | -6.86*** | 1.99 |

Table 2: Regression coefficients for the model predicting reference sentence as DO-fronted non-canonical (DOSV) against variant as canonical (SDOV) ($N = 1663$; all significant predictors denoted by *** $p < 0.001$; VIF = Variance Inflation Factor)

found no such influence of dependency length in IO-fronted sentences. In the present work, we find that while OSV sentences optimize for surprisal minimization generally, only DO-fronted sentences tend to minimize dependency length.

Classification Analysis

In this section, we examine the impact of each factor in classifying whether a sentence is a corpus reference sentence or a variant. We use 10-fold cross-validation of the entire dataset (72833 data points) to evaluate the classifier’s performance. Table 4 presents the prediction accuracy of our various models across three constructions *viz.*, object-fronted (OSV), DO-fronted (DOSV), and IO-fronted (IOSV) in our dataset.

In terms of individual performance (top block in Table 4), adaptive LSTM and trigram surprisal achieved the highest classification accuracy across all three constructions. PCFG surprisal is the second best predictor followed by dependency length and IS score in decreasing order for OSV and IOSV constructions. The poor performance of PCFG surprisal in predicting DOSV sentences is mainly due to its bias towards frequent syntactic structures, in contrast to dependency length, which quantifies memory load within sentence. These findings suggest that a syntactically complex sentence can be less taxing if the processing is facilitated via DLM or probable word sequences in the sentence.

Over a baseline model containing adaptive LSTM and trigram surprisal measures (bottom block in Table 4), none of the predictors—PCFG surprisal, dependency length, and IS score—induced any increase in the classification performance for OSV (second column) and IOSV constructions (last column). However, for DOSV construction, PCFG surprisal significantly helped classify reference sentences above and beyond the baseline model ($p < 0.001$ using McNemar’s two-tailed significance test). Dependency length failed to contribute beyond the model containing all the surprisal measures. Interestingly, IS score successfully predicted DO-fronted constructions beyond the model containing all predictors, implying that fronting direct objects is driven by considerations of givenness in discourse. In another experiment, we also tested whether dependency length significantly contributes beyond a model with only adaptive LSTM surprisal, but found no significant impact in predicting corpus sentences across the three constructions. This implies that the impact of DLM is entirely subsumed by adaptive LSTM surprisal.

| Predictors | Estimate | Std. Error | Z-value | VIF |
|-------------------------|--------------|------------|-----------|------|
| Intercept | 0.20 | 0.109 | 1.875 | – |
| Dependency length | 0.15 | 0.142 | 1.045 | 1.08 |
| IS score | 0.13 | 0.074 | 1.767 | 1.10 |
| PCFG surprisal | -0.89 | 0.238 | -3.725*** | 1.08 |
| Trigram surprisal | -3.05 | 0.382 | -7.976*** | 1.34 |
| Adaptive LSTM surprisal | -2.83 | 0.365 | -7.78*** | 1.43 |

Table 3: Regression coefficients for the model predicting reference sentence as IO-fronted non-canonical (IOSV) against variant as canonical (SIOV) ($N = 1353$; all significant predictors denoted by *** $p < 0.001$; VIF = Variance Inflation Factor)

In sum, our classification analyses involving 10-fold cross-validation, focused on *prediction*, stand in contrast to our regression analyses, which involved *fitting* the data. Our classification findings refute the hypothesis that DLM significantly predicts non-canonical ordering choices, once expectation-based factors and discourse considerations are accounted for. Instead, surprisal minimization, captured by inter- and intra-sentential language models like adaptive LSTM and trigram surprisal, emerges as the primary driver of syntactic ordering choices in Hindi. Our work extends beyond prior research that solely incorporates intra-sentential surprisal measures for modeling syntactic choices (Rajkumar et al., 2016; Ranjan, Agarwal, & Rajkumar, 2019), and aligns with the recent findings that have attested the role of discourse context in sentence processing (Qian & Jaeger, 2012; van Schijndel & Linzen, 2018; Ranjan, van Schijndel, et al., 2022).

Human Evaluation

While our previous analyses relied entirely on corpus occurrences as proxies for human preference data, this section investigates how well our classification model simulates real human judgments from 44 native speakers. As described in the Methods section, we collected human judgments using 164 randomly chosen reference-variant pairs from our full dataset (72833). We employed 10-fold cross-validation on the full dataset to assess our classifier’s performance on these 164 pairs, trained with all features. Subsequently, we compared the classifier’s output with the real-human choices. Table 5 reports our human evaluation outcomes.

Our classifier consistently demonstrated better accuracy in predicting the human-preferred constructions across all the items we tested than in predicting original corpus choices. The increased accuracy for human labels (70%) over corpus labels (45%) in DO-fronted constructions revealed that native speakers generally prefer canonical forms over non-canonical ones even though the non-canonical sentences minimized dependency length on average (see Figure 1 and Table 2), and our model is better at picking up those instances. We also found that Hindi participants favored IO-fronted constructions (71.43%) over DO-fronted constructions (55%) as captured by the human-corpus agreement accuracy, which reflects the number of sentences where the corpus reference sentence was preferred by human raters. Further data analysis revealed that our model struggled to predict human la-

| Predictor(s) | (Ref: OSV vs Var: SOV) | (Ref: DOSV vs Var: SDOV) | (Ref: IOSV vs Var: SIOV) |
|---|------------------------|--------------------------|--------------------------|
| IS Score | 52.19 | 53.88 | 50.92 |
| Dependency length | 65.24*** | 68.49*** | 58.91*** |
| PCFG surprisal | 67.22 | 59.05*** | 75.91*** |
| Trigram surprisal | 83.16*** | 78.95*** | 87.29*** |
| Adaptive LSTM surprisal | 83.84 | 79.98 | 88.32 |
| Baseline = Adaptive LSTM + trigram surprisal | 85.18*** | 81.24** | 89.43* |
| Baseline + PCFG surprisal | 84.72 | 80.04*** | 89.95 |
| Baseline + PCFG surprisal + dependency length | 84.65 | 79.86 | 89.95 |
| All predictors | 85.04 | 80.46* | 90.02 |

Table 4: Individual and collective prediction accuracies (Random accuracy = 50%; *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ McNemar’s two-tailed significance compared to model on previous row; Refer to Table 1 for our dataset distribution)

| Dataset Type (#) | Corpus labels Prediction Accuracy (%) | Human labels Prediction Accuracy (%) | Human-Corpus Agreement (%) |
|-------------------------|---------------------------------------|--------------------------------------|----------------------------|
| Ref: OSV; Var: SOV (41) | 53.66 | 70.73 | 63.41 |
| DO-fronted (20) | 45.00 | 70.00 | 55.00 |
| IO-fronted (21) | 61.90 | 71.43 | 71.43 |
| Full dataset (164) | 76.65 | 79.04 | 85.63 |

Table 5: Model’s performance on original corpus labels and the human labels from 44 native Hindi speakers

bels when participants themselves disagree among each other. We observed significant disagreement among native speakers for non-canonical (OSV) orders, and in particular, for DO-fronted reference sentences, thus serving as a primary reason underlying the differing performance of our classifier in predicting human and corpus labels.

Discussion

Dependency Length Minimization seems to be a general property of many natural languages and it does seem to significantly predict general word ordering in Hindi. However, our work indicates that discourse expectations, as captured by adaptive LSTM surprisal and givenness seem to be the primary influences on non-canonical sentence production. Notably, discourse-enhanced surprisal entirely subsumes the impact of dependency length minimization effects in predicting Hindi syntactic choices. It is possible that dependency length partially helps determine when non-canonical constructions occur, but our results indicate that at least some other factors are also at work. We suspect that other cognitive mechanisms underlying surprisal minimization include accessibility of multi-word sequences (Reali & Christiansen, 2007), considerations of local coherence effects (Tabor, Galantucci, & Richardson, 2004), and discourse adaptation, incorporating both lexical and syntactic priming (Gries, 2005; Tooley & Traxler, 2010; Reitter, Keller, & Moore, 2011; Ranjan, van Schijndel, Agarwal, & Rajkumar, 2022).

Previous studies in Hindi have noted a weak locality effect, affecting sentences that are either syntactically complex or that involve rare constructions (Vasishth & Lewis, 2006; Ranjan, Rajkumar, & Agarwal, 2022). However, our present findings reveal that, with the inclusion of discourse factors, the locality effect ceases to be a general predictor, even in less common non-canonical sentences. The present findings, therefore, align with the idea in the literature that expectation effects are likely to be prevalent in SOV languages, whereas

locality effects are observed in SVO languages (Vasishth, Suckow, Lewis, & Kern, 2010; Ferrer-i-Cancho, 2015, 2017)

A recent study by Hahn, Jurafsky, and Futrell (2020), involving an extensive analysis of 51 languages, demonstrated that dependency-length minimization could be an epiphenomenon, sufficiently accounted for by optimizing the general predictability and parseability of sentences without requiring any additional constraints. Similarly, Engelmann, Jager, and Vasishth (2019) noted that decay component (representative of locality) in Lewis and Vasishth (2005)’s model did not improve fit in their simulation data across various dependency types. These findings possibly suggest that *decay*, the underlying cognitive construct behind “locality” may lack strong empirical support (Oberauer & Lewandowsky, 2013, 2014; Stone, von der Malsburg, & Vasishth, 2020; Jager, Merten, Van Dyke, & Vasishth, 2020), although see Hardt, Nader, and Nadel (2013). This casts doubt on the cause of dependency length minimization being solely attributed to decay in memory, and proposes an alternative explanation where locality might be reducible to a more general memory interference effect (Vasishth et al., 2012; Ranjan, Agarwal, & Rajkumar, 2024). Future work needs to investigate the role of interference while controlling for locality and surprisal in syntactic choice phenomena in Hindi.

Our results also have implications for the Information Locality Hypothesis (ILH), which integrates lossy context surprisal and mutual information (Futrell, 2019; Futrell, Gibson, & Levy, 2020). ILH suggests that words predicting each other should exhibit locality effects, and processing difficulty is related to a word’s expected log probability given a noisy memory representation. With this view, ILH, thus is thought to unify surprisal and dependency locality, and proposes a metric based on point-wise mutual information (PMI) between word pairs. Our work uses an LSTM language model with an unbounded context, which effectively captures and subsumes the effects of locality for syntactic choice. Thus offering some preliminary evidence for ILH. Nevertheless, further research is required to test the original formulation of ILH using PMI metric for syntactic choice in Hindi.

Overall, our results provide converging evidence that while speakers of Hindi, an SOV language, display a weak tendency to minimize dependency length, the preverbal constituent orderings are primarily shaped by surprisal minimization leveraging both intra- and inter-level context information.

Acknowledgments

We would like to express our gratitude to the anonymous reviewers of HSP-2023 and CogSci-2024 for their invaluable feedback. The human data in this paper was collected according to Cornell IRB protocol IRB0010461 (2107010461).

References

- Bader, M., & Meng, M. (2023). Processing noncanonical sentences: Effects of context on online processing and (mis) interpretation. *Glossa Psycholinguistics*, 2(1).
- Baker, P., Hardie, A., McEnery, T., Cunningham, H., & Gaizauskas, R. (2002). Emille: a 67-million word corpus of indic languages: data collection, mark-up and harmonization. In *Proceedings of Irec 2002* (pp. 819–827). Lancaster University.
- Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., & Xia, F. (2009). A multi-representational and multi-layered treebank for Hindi/urdu. In *Proceedings of the third linguistic annotation workshop* (pp. 186–189). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1698381.1698417>
- Bock, J. K. (1982). Towards a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 1–47.
- Chomsky, N. (1965). *Aspects of the theory of syntax* (Vol. 11). The MIT press.
- Clark, H. H., & Haviland, S. E. (1977). Comprehension and the Given-New Contract. In R. O. Freedle (Ed.), *Discourse Production and Comprehension* (pp. 1–40). Hillsdale, N. J.: Ablex Publishing.
- Comrie, B. (1981). *Language universals and linguistic typology: syntax and morphology / bernard comrie*. Blackwell Oxford.
- Dammalapati, S., Rajkumar, R., Ranjan, S., & Agarwal, S. (2021). Effects of duration, locality, and surprisal in speech disfluency prediction in english spontaneous speech. In *Proceedings of the society for computation in linguistics* (Vol. 4, p. 10). Retrieved from <https://scholarworks.umass.edu/scil/vol4/iss1/10/>
- Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2), 193–210. Retrieved from http://scholar.google.com/scholar.bib?q=info:lulLoWII1DoJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=0
- Engelmann, F., Jager, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, 43(12), e12800. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12800> doi: 10.1111/cogs.12800
- Ferreira, V. S. (1996). Is it better to give than to donate? syntactic flexibility in language production. *Journal of memory and language*, 35(5), 724–755.
- Ferrer-i-Cancho, R. (2015). The placement of the head that minimizes online memory. *Language Dynamics and Change*, 5(1), 114–137. Retrieved from <http://booksandjournals.brillonline.com/content/journals/10.1163/22105832-00501007> doi: <https://doi.org/10.1163/22105832-00501007>
- Ferrer-i-Cancho, R. (2017, 05). The placement of the head that maximizes predictability. an information theoretic approach. *Glottometrics*, 39, 38–71.
- Futrell, R. (2019, August). Information-theoretic locality properties of natural language. In *Proceedings of the first workshop on quantitative syntax (quasy, syntaxfest 2019)* (pp. 2–15). Paris, France: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-7902> doi: 10.18653/v1/W19-7902
- Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3), e12814. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12814> doi: 10.1111/cogs.12814
- Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, 96(2), 371–412.
- Gambhir, V. (1981). *Syntactic restrictions and discourse functions of word order in standard Hindi*. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1–76.
- Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O’Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium*. Cambridge, MA: MIT Press. Retrieved from <http://www.ling.uni-potsdam.de/~vasishth/Papers/Gibson-Cognition2000.pdf>
- Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(4), 365–399. Retrieved from <http://dx.doi.org/10.1007/s10936-005-6139-3> doi: 10.1007/s10936-005-6139-3
- Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5), 2347–2353. Retrieved from <https://www.pnas.org/content/117/5/2347> doi: 10.1073/pnas.1910923117
- Hale, J. (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* (pp. 1–8). Pittsburgh, Pennsylvania: Association for Computational Linguistics. Retrieved from <http://dx.doi.org/10.3115/1073336.1073357> doi: 10.3115/1073336.1073357
- Hale, J. (2006). Uncertainty about the rest of the sentence.

- Cognitive science*, 30(4), 643–672.
- Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: the role of active forgetting in memory. *Trends in cognitive sciences*, 17(3), 111–120.
- Hawkins, J. A. (1994). *A Performance theory of order and constituency*. New York: Cambridge University Press.
- Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.
- Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press.
- Hörnig, R., Oberauer, K., & Weidenfeld, A. (2005). Two principles of premise integration in spatial reasoning. *Memory & Cognition*, 33(1), 131–139.
- Jager, L. A., Merten, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111, 104063. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0749596X19300956> doi: <https://doi.org/10.1016/j.jml.2019.104063>
- Joachims, T. (2002). Optimizing search engines using click-through data. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 133–142). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/775047.775067> doi: 10.1145/775047.775067
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, 20(2), 137–194.
- Kachru, Y. (1982). Conjunct verbs in hindi-urdu and persian. *South Asian Review*, 6(3), 117–126. Retrieved from <https://doi.org/10.1080/02759527.1982.11933096> doi: 10.1080/02759527.1982.11933096
- Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94(2), 113–147.
- Kristensen, L. B., Engberg-Pedersen, E., & Poulsen, M. (2014). Context improves comprehension of fronted objects. *Journal of Psycholinguistic Research*, 43, 125–140.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126 - 1177. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0010027707001436> doi: <http://dx.doi.org/10.1016/j.cognition.2007.05.006>
- Lewis, R. L., & Vasishth, S. (2005, May). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29, 1–45. Retrieved from <http://www.ling.uni-potsdam.de/~vasishth/Papers/Lewis-VasishthCogSci2005.pdf>
- Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159–191. Retrieved from <http://www.lingviko.net/JCS.pdf>
- Mohan, K., & Mohan, T. (1994). Issues in word order in south asian languages: Enriched phrase structure or multidimensionality? In M. Butt, T. H. King, & G. Ramchand (Eds.), *Theoretical perspectives on word order in south asian languages* (p. 153-184). Stanford, CA: Center for the Study of Language and Information.
- Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General*, 142(2), 380–411. doi: 10.1037/a0029588
- Oberauer, K., & Lewandowsky, S. (2014). Further evidence against decay in working memory. *Journal of Memory and Language*, 73, 15–30.
- Qian, T., & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive Science*, 36(7), 1312–1336. Retrieved from <http://dx.doi.org/10.1111/j.1551-6709.2012.01256.x> doi: 10.1111/j.1551-6709.2012.01256.x
- Rajkumar, R., van Schijndel, M., White, M., & Schuler, W. (2016). Investigating locality effects and surprisal in written english syntactic choice phenomena. *Cognition*, 155, 204–232. Retrieved from <http://www.sciencedirect.com/science/article/pii/S001002771630155X/>
- Rajkumar, R., & White, M. (2014). Better surface realization through psycholinguistics. *Language and Linguistics Compass*, 8(10), 428–448. Retrieved from <http://dx.doi.org/10.1111/lnc3.12090> (ISSN: 1749-818X) doi: 10.1111/lnc3.12090
- Ranjan, S., Agarwal, S., & Rajkumar, R. (2019, June). Surprisal and interference effects of case markers in Hindi word order. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 30–42). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-2904>
- Ranjan, S., Agarwal, S., & Rajkumar, R. (2024). Interference predicts locality: Evidence from an SOV language. In *Proceedings of the Society for Computation in Linguistics (SCiL)*.
- Ranjan, S., Rajkumar, R., & Agarwal, S. (2022). Locality and expectation effects in hindi preverbal constituent ordering. *Cognition*, 223, 104959. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0010027721003826> doi: <https://doi.org/10.1016/j.cognition.2021.104959>
- Ranjan, S., van Schijndel, M., Agarwal, S., & Rajkumar, R. (2022). Dual mechanism priming effects in hindi word order. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*. Retrieved from <https://aclanthology.org/2022.aacl-main.68/>
- Ranjan, S., van Schijndel, M., Agarwal, S., & Rajkumar, R. (2022, December). Discourse context predictability effects in Hindi word order. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10390–10406). Abu Dhabi, United Arab Emirates: As-

- sociation for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.emnlp-main.710> doi: 10.18653/v1/2022.emnlp-main.710
- Ranjan, S., & von der Malsburg, T. (2023). A bounded rationality account of dependency length minimization in hindi. In M. Goldwater, F. Anggoro, B. Hayes, & D. Ong (Eds.), *Proceedings of the 45th annual meeting of the cognitive science society*. Sidney, Australia: Cognitive Science Society. Retrieved from <https://escholarship.org/uc/item/85q7x1nd>
- Ranjan, S., & von der Malsburg, T. (2024). Work smarter...not harder: Efficient minimization of dependency length in SOV languages. In L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th annual meeting of the cognitive science society*. Rotterdam, Netherlands: Cognitive Science Society. Retrieved from <https://arxiv.org/abs/2404.18684>
- Reali, F., & Christiansen, M. (2007, 03). Word chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly journal of experimental psychology (2006)*, 60, 161-70. doi: 10.1080/17470210600971469
- Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, 35(4), 587–637. Retrieved from <http://www.david-reitter.com/pub/reitter2011syntacticpriming.pdf> doi: 10.1111/j.1551-6709.2010.01165.x
- Scontras, G., Badecker, W., Shank, L., Lim, E., & Fedorenko, E. (2014). Syntactic complexity effects in sentence production. *Cognitive Science*, 39(3), 559–583. Retrieved from <http://dx.doi.org/10.1111/cogs.12168> doi: 10.1111/cogs.12168
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27.
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.
- Stolcke, A. (2002). SRILM — An extensible language modeling toolkit. In *Proc. icslp-02*.
- Stone, K., von der Malsburg, T., & Vasishth, S. (2020). The effect of decay and lexical uncertainty on processing long-distance dependencies in reading. *PeerJ*, 8, e10438.
- Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4), 355 - 370. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0749596X04000087> doi: <https://doi.org/10.1016/j.jml.2004.01.001>
- Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, 105(2), 300–333. Retrieved from <http://www.sciencedirect.com/science/article/B6T24-4M7CDMS-2/2/e095449f6439b30003822a5838e53786> doi: DOI:10.1016/j.cognition.2006.09.011
- Tooley, K. M., & Traxler, M. J. (2010). Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, 4(10), 925–937.
- van Schijndel, M., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3), 522–540.
- van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4704–4710).
- Vasishth, S. (2004). Discourse context and word order preferences in Hindi. *Yearbook of South Asian Languages*, 113–127. Retrieved from <http://www.ling.uni-potsdam.de/~vasishth/pdfs/Vasishth2004YearbookSAL.pdf>
- Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4), 767–794. Retrieved from <http://www.ling.uni-potsdam.de/~vasishth/Papers/Vasishth-Lewis-Language2006.pdf>
- Vasishth, S., Shaher, R., & Srinivasan, N. (2012). The role of clefting, word order and given-new ordering in sentence comprehension: Evidence from Hindi. *Journal of South Asian Linguistics*, 5, 35-56. Retrieved from <http://tiger.sprachwiss.uni-konstanz.de/~jsal/ojs/index.php/jsal/article/viewFile/27/26>
- Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4), 533-567. Retrieved from <https://doi.org/10.1080/01690960903310587> doi: 10.1080/01690960903310587
- Yadav, H., Vaidya, A., & Husain, S. (2017). Keeping it simple: Generating phrase structure trees from a Hindi dependency treebank. In *Tlt*.