Clause Final Verb Prediction in Hindi: Evidence for Noisy Channel Model of Communication

Kartik Sharma[†*], **Niyati Bafna**[‡*] **and Samar Husain**[†]
[†]Indian Institute of Technology Delhi, [‡]Charles University, Faculty of Mathematics and Physics
kartik.sharma.cs117@cse.iitd.ac.in, 64780815@o365.cuni.cz,
samar@iitd.ac.in

Abstract

Verbal prediction has been shown to be critical during online comprehension of Subject-Object-Verb (SOV) languages. In this work we present three computational models to predict clause final verbs in Hindi given its prior arguments. The models differ in their use of prior context during the prediction process – the context is either noisy or noise-free. Model predictions are compared with the sentence completion data obtained from Hindi native speakers. Results show that models that assume noisy context outperform the noise-free model. In particular, a lossy context model that assumes prior context to be affected by predictability and recency captures the distribution of the predicted verb class and error sources best. The success of the predictability-recency lossy context model is consistent with the *noisy channel hypothesis* for sentence comprehension and supports the idea that the reconstruction of the context during prediction is driven by prior linguistic exposure. These results also shed light on the nature of the noise that affects the reconstruction process. Overall the results pose a challenge to the *adaptability hypothesis* that assumes use of noise-free preverbal context for robust verbal prediction.

1 Introduction

Research on sentence comprehension has conclusively established the widespread role of prediction during online processing (e.g., Marslen-Wilson, 1973; Altmann and Kamide, 1999; Staub and Clifton, 2006; Kutas and Hillyard, 1984). It is known that comprehenders actively anticipate the upcoming linguistic material prior to receiving that information during listening or reading (Luke and Christianson, 2016; Staub, 2015). The role of active prediction during comprehension has particularly been emphasized for processing of SOV languages (e.g., Konieczny, 2000; Yamashita, 1997; Friederici and Frisch, 2000). In particular, it has been argued that preverbal nominal features such as case-markers are effectively used to make precise prediction regarding the clause final verb. Indeed, the ADAPTABILITY HYPOTHESIS states that owing to the typological properties, the prediction system in SOV languages is particularly *adapted* to make effective use of preverbal linguistic material to make robust clause final verbal prediction (Vasishth et al., 2010; Levy and Keller, 2013). Evidence for the *adaptability hypothesis* come from various behavioral experiments that show effective use of case-markers to make clause final verbal prediction (e.g., Husain et al., 2014), facilitation at the verb when the distance between the verb and its prior dependent increase (e.g., Konieczny, 2000), and lack of structural forgetting in the face of complex linguistic environment (e.g., Vasishth et al., 2010). On the other hand, the NOISY CHANNEL HYPOTHESIS assumes that prediction during comprehension is required to accommodate uncertainty in the input (Gibson et al., 2013; Kurumada and Jaeger, 2015). In other words, the hypothesis posits that comprehenders have the knowledge that speakers make mistakes during production, hence, comprehenders need to reconstruct the received input (Ferreira and Patson, 2007).

The two hypotheses stated above make distinct assumptions regarding the utilization of pre-verbal context towards making clause final verbal predictions in SOV languages. One way to operationalize the predictions of the *adaptability hypothesis* is to assume that the preverbal linguistic material will be faithfully used to make verbal prediction, the *noisy channel hypothesis* on the other hand, assumes that the preverbal context is noisy and therefore subject to reconstruction. One consequence of this would be that the *adaptability hypothesis* would predict that verbal prediction should be robust while the *noisy channel hypothesis* would predict that verbal prediction should be susceptible to errors. In

*Equal contribution by KS and NB.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 160–170

addition, the two hypotheses would make distinct prediction regarding the nature of errors that might occur during clause final verbal prediction.

In order to probe the two hypotheses stated earlier, in this work, we investigate various incremental models that use local linguistic features to predict clause final verbal prediction in Hindi (an SOV language). The distribution of these model predictions is compared with human data. In particular, we investigate to what extent the models are able to capture the nature of both grammatical as well as ungrammatical verbal predictions when compared to data collected from native speakers of Hindi. Further, in order to probe the assumptions of the *noisy channel hypothesis* more closely, we probe multiple noise functions to investigate the nature of preverbal context reconstruction during prediction.

The paper is arranged as follow, in Section 2 we briefly describe the experimental results that we model. Section 3 provide the necessary details regarding methodology (data/tools, model evaluation, etc.). In Sections 4 and 5 we respectively discuss the n-gram surprisal and the lossy-surprisal models. Section 6 presents the results. Section 7 discusses the current findings and its implications. We conclude the paper in Section 8.

2 Background

In spite of the proposed central role of verb prediction during online processing of Hindi (e.g., Vasishth and Lewis, 2006; Agrawal et al., 2017; Husain et al., 2014), there is a surprising lack of any modeling attempt to understand the processes that subserve verbal predictions in the language. While there are computational metrics that model reading time data (e.g., Hale, 2001; Shain et al., 2016; Futrell et al., 2020), a computational model that makes precise verbal prediction in SOV languages has not been investigated thoroughly (but see, Grissom II et al., 2016, for an initial attempt). Understanding the mechanisms that subserve verbal prediction in SOV languages is critical to understanding how these languages are processed (cf. Konieczny, 2000; Vasishth et al., 2010; Husain et al., 2014; Levy and Keller, 2013; Kuperberg and Jaeger, 2016). Our work fills this gap in the literature. In this section we summarize the key results of a recent study by Apurva and Husain (2020) who investigated the nature of verbal prediction in Hindi using a series sentence completion studies (Staub et al., 2015). Later, in sections 4, 5 we present three computational models to account for these results.

2.1 Completion Study Results

Apurva and Husain (2020) used the sentence completion paradigm (Taylor, 1953) to probe the nature of clause final verbal prediction when differing the number of preverbal nouns that precede the to-be-completed target verb. The number of nouns ranged from 1 to 3 and appeared in different case-marker order. All preverbal nouns were proper nouns. Example 1 shows some of the conditions where 3 preverbal nouns preceded the target verb. In the example, *ne* is the Ergative case-marker, *ko* is the Accusative case-marker and *se* is the Ablative case-marker. In all, there were 6 conditions in this experiment (ne-ko-se, ne-se-ko, ko-ne-se, ko-se-ne, se-ko-ne, se-ne-ko). 36 native speakers participated in the 3-NP condition experiments. Similar to the 3-NP conditions, the 1-NP and 2-NP items had proper nouns and the nouns occurred in various case-marker order. 25 native speakers participated in the 1-NP and 2-NP condition experiments.

(1) a. ne-ko-se
pooja-**ne** urmila-**ko** suneet-**se** ...
Pooja-ERG Urmila-ACC Suneet-ABL ...

b. ne-se-ko
pooja-**ne** urmila-**se** suneet-**ko** ...
Pooja-ERG Urmila-ABL Suneet-ACC ...

The key result from these completion studies was that the number of ungrammatical verbal completions increased as the number of preverbal nominals increased. For the 1-NP conditions the percentage ungrammatical completions was 4%, for the 2-NP conditions this was 8%, while for the 3-NP conditions the ungrammatical completions increased to 15%.

In addition, the completion data was also analyzed for the nature of grammatical and ungrammatical verbal completions. Completions were analyzed based on the verb classes rather than lexical identity (cf. Luke and Christianson, 2016). The data contains a distribution over a total of 18 verbs classes for the 2-NP and 3-NP conditions. In majority of the grammatical completions, Hindi native speakers posit simple syntactic structures (in terms of the number of clausal embeddings and the number of core argument structure). For the 2-NP conditions, the topmost verb classes were *T* (Transitive verb), *IN* (Intransitive verb), and *DT* (Ditransitive verb). For the 3-NP conditions, *CAUS* (Causative verb) and *T DT* (Transitive non-finite

verb followed by a ditransitive matrix verb) were consistently the most frequent, covering at least 50% of completions between them for all conditions. Some of the other classes observed were *DT*, *N T DT*, and *DT DT*. Interestingly, while the 3-NP conditions can be grammatically completed using a double embedded structure (e.g., IN DT DT), such cases were not found in the completion data.

Among the ungrammatical verb completions across various conditions, *N DT*, *IN DT* and *CAUS* were consistently the most frequent verb classes predicted. Similar to the trend in the grammatical completions discussed above, the parser posits simple structures even when making mistakes. Additionally, a closer analysis of the ungrammatical completions showed the formation of *locally coherent* parses (Tabor et al., 2004) for the various 3-NP conditions where the first noun was ignored and only the 2nd and the 3rd nouns were used to make the prediction (we call these N2-N3 errors). Other errors were made when either N2 or N3 were ignored to make the prediction (we call these N1-N3, N1-N2 errors respectively). The errors also show a subject primacy effect (Häussler and Bader, 2015; Knoedler et al., 1999) where the presence of an Ergative case marker on N1 is not forgotten. This leads to lack of passive predictions in such cases.[2]

To sum up, the key results of the completion studies were, (a) verb prediction was good in 1-NP and 2-NP conditions, (b) predictions deteriorated in 3-NP conditions, (c) grammatical verbal completions are syntactically simple rather than complex (e.g., clausal embeddings are avoided), (d) error types for the 3-NP conditions show use of two preverbal NPs to make predictions, as well as being sensitive to subject primacy.

Table 1 provides the details on the number of grammatical and ungrammatical completions over all conditions. Also see Table 3 for verb class numbers for the 2-NP conditions. Table 2 shows examples of various error types in the 3-NP conditions.

3 Methodology

3.1 Data and Tools

We use the monolingual Hindi corpus developed by IIT Bombay (Kunchukuttan et al., 2017). It is a collection of raw sentences of Hindi taken from various sources (HindMonoCorp (Bojar et al., 2014), BBC, Wikipedia etc.). For training our models, we use the first 5 million sentences of this data. For the sentence simplification step (described in the Section 3.2), we use the ISC dependency parser for Hindi.[3] Moreover, as the sentence completion experiment included only animate nouns in various items (see Section 2), we use an additional animacy annotation (Jena et al., 2013) to label the nouns accordingly.

3.2 Sentence Simplification

A key aim of the behavioral experiments discussed in Section 2 was to investigate the role of preverbal arguments on clause final verbal prediction. Consequently, our models had to be trained on sentences with various features (e.g., case-marker, animacy) of the preverbal arguments . Since the raw data may contain other intervening material (nominal modifiers, verbal adjuncts, etc.),[4] the task necessitated removal of such material from the training corpus to render it more tractable to the appropriate computational model. Thus, we simplify each sentence in the training data by removing these intervening materials while ensuring that the grammaticality of the sentence remains intact.[5] This, of course, implies that the model only uses the local argument structure to make the necessary verbal prediction.

The sentence simplification process preserves verbal and nominal arguments, such as direct/oblique objects, case-markers, and auxiliaries, but removes adjective phrases, relative clauses, and adjuncts. It treats conjunct structures as separate components. It identifies intra-sentential noun ellipsis and truncates a sequence that displays such a structure, while processing its other verbs. For example:

police-**ne** giraftari warrant
Police-ERG arrest warrant
milne-ke baad somwar raat-ko
get-INF-ACC-GEN after Monday night
Ratan-ke vakeel-**se**
Ratan-GEN-ACC lawyer-ABL

[2] See Sections 1 and 2 of the supplementary material for additional details regarding the word order in Hindi, experimental conditions, predicted verb classes predicted and examples of various errors during the completion study.

[3] https://bitbucket.org/account/user/iscnlp/projects/ISCNLP. It is an implementation of the incremental transition-based arc-eager parsing algorithm (Nivre, 2008). The parser is trained on the Hyderabad Dependency Treebank (Bhatt et al., 2009) and is reported to have a UAS of 93.52% and an LAS of 87.77% (Bhat, 2017)

[4] Refer to Section 3 of the supplementary material for statistics on the same.

[5] Additional details regarding procedure and testing have been provided in Section 4 of the supplementary material.

Verb Class	Grammatical completions	Ungrammatical completions
T+DT	170	6
CAUS	140	9
N+DT+DT	51	1
DT+DT	24	0
T+T	21	0
N+CAUS	19	0
N+T+DT	17	5
DT	10	1
CAUS+T	5	0
CAUS+DT	4	0
IN+DT	2	8
T	2	5
N+DT	2	15
N+T	1	2
DT+IN	0	1
DT+T	0	1
IN	0	1
IN+DT	0	2
Other	7	4

Table 1: Grammatical and ungrammatical completions across all 2-NP and 3-NP conditions. *v1+v2* signifies an embedded structure with *v1* as the embedded non-finite verb and *v2* as the matrix verb. In the case of *n+v1+v2*, *n* is part of the *v1* non-finite clause and *v2* is the matrix verb. IN: Intransitive, CAUS: Causative, T: Transitive, DT: Ditransitive, N: Noun.

Error type	Example
N1 N2	N1-ne N2-ko N3-se peeta tha 'hit PAST'
N1 N3	N1-ne N2-ko N3-se kuchaa mangaa 'something asked'
N2 N3	N1-se N2-ne N3-ko introduce kiya 'introduce do'

Table 2: Sample completions for various error types in some 3-NP conditions. Completions are underlined. Note: the completions are grammatical if we ignore the striked-out phrase; else they are ungrammatical. ne=Ergative case-marker, ko=Accusative case-marker, se=Ablative case-marker.

sampark-kiya-tha
communicate-P.Perf

↓

police-**ne** vakeel-**se** sampark-kiya-tha
Police-ERG lawyer-ABL communicate-P.Perf

We also flatten all the nouns in the data to "noun tokens" by merging the noun and its corresponding case-marker. Since we are interested in capturing the variations of the completions for different order of case-markers in the prompt, we can abstract away from the lexicality of the nouns. Thus, we replace the nominal lexical item with its corresponding label depending on whether it is animate (**A**) or not (**N**). Such an abstraction is well motivated considering that humans are known to be sensitive to both syntactic part-of-speech tags as well as lexical semantics during sentence processing (e.g., Demberg and Keller, 2008; Trueswell et al., 1994).

3.3 Experiment Design

Given the abstract nominals and their case-marker, a model's task is to complete the input string with an appropriate verb phrase. For example, if the model is given 3 noun tokens (each with a unique case-marker) with the lexical item replaced with a label **A** denoting *animate*, the task is to predict a verb phrase from this context. End of prediction is signalled as a punctuation.

We note that, given a context, the model makes the prediction in an incremental fashion, rather than producing a one-shot phrase. This means that once a word is predicted, the model considers it as part of the context for the prediction of the next word. For example, given "**A**-ne **A**-ko **A**-se", the model completes the sentence with $w_1 w_2 w_3$ in the following manner:

A-ne **A**-ko **A**-se $\Rightarrow w_1$

A-ne **A**-ko **A**-se $w_1 \Rightarrow w_2$

A-ne **A**-ko **A**-se $w_1\ w_2 \Rightarrow w_3$

All implemented models discussed in Section 4 and Section 5, use the 1/2/3 preverbal arguments as context. The rationale for use of local context is driven by the goal to model the role of argument structure in verbal prediction (see Section 2). Interestingly, the automatically parsed Hindi corpus (Bojar et al., 2014) shows that arguments (when compared to adjuncts) tend to be closer to the verb[6] suggesting that the critical information needed to

[6] Arguments are at an average distance of 3.8 from the verb while adjuncts have mean dependency distance of 4.5.

Cond	M_Vc	Total	VC count	Cond	M_Vc	Total	VC count	Cond	M_Vc	Total	VC count
n1c1n2c1	COP	1	3	n1c2n2c4	DT	8	23	n1c3n2c4	T	71	84
n1c1n2c1	T	2	3	n1c2n2c4	T	14	23	n1c4n2c1	IN	10	23
n1c1n2c3	DT	3	22	n1c3n2c1	DT	3	22	n1c4n2c1	T	13	23
n1c1n2c3	T	19	22	n1c3n2c1	T	19	22	n1c4n2c2	CAUS	8	82
n1c1n2c4	COP	4	23	n1c3n2c2	DT	30	89	n1c4n2c2	DT	32	82
n1c1n2c4	EXP	2	23	n1c3n2c2	T	59	89	n1c4n2c2	IN	2	82
n1c1n2c4	IN	3	23	n1c3n2c3	DT	2	6	n1c4n2c2	T	40	82
n1c1n2c4	T	14	23	n1c3n2c3	T	4	6	n1c4n2c3	CAUS	2	77
n1c2n2c1	DT	4	21	n1c3n2c4	CAUS	1	84	n1c4n2c3	DT	5	77
n1c2n2c1	T	17	21	n1c3n2c4	COP	1	84	n1c4n2c3	T	70	77
n1c2n2c3	DT	6	24	n1c3n2c4	DT	1	84	n1c4n2c4	T	1	1
n1c2n2c3	T	18	24	n1c3n2c4	EXP	5	84				
n1c2n2c4	CAUS	1	23	n1c3n2c4	IN	5	84				

Table 3: 2-NP Predictions: c1=Nom, c2=Erg, c3=Acc, c4=Abl; IN: Intransitive, CAUS: Causative, T: Transitive, DT: Ditransitive, N: Noun. 'Total' refers to the number of instances of the condition, 'VC count' refers to the number of instances of the corresponding verb class.

predict the verb should be accessible locally. In addition, we place an upper limit on the no. of predicted words – 2 words for 2-NP conditions and 3 for 3-NP.[7] Given the cognitive validity of limited beam-size (e.g., Boston et al., 2011), we only pick the top 50 predictions for further analyses.

Both human and model completions are manually annotated with verb classes based on the valency of the predicted verb. In addition, any nominal argument prediction was also annotated. Verb classes were labeled as *IN* (intransitive), *T* (transitive), *DT* (ditransitive), *CAUS* (causative), or combinations of the above in case a combination of non-finite and matrix verbs is predicted.

For example, the following phrase contains a transitive verb preceded by its object noun:

(2) khaana khaaya ⟶ N T
food eat-PT

Verb classes are used for comparing model output with human data as predictions are known to be graded rather than all-or-nothing lexical prediction (Luke and Christianson, 2016; Staub, 2015). Additionally, we don't predict the verb classes directly to keep the model output consistent with the human data. These completions are then labelled for grammaticality automatically; given the prompt condition and the verb class of the completion, we can infer the grammaticality of the sentence.[8]

[7]No significant change in the set of predictions was observed on increasing these numbers any further.

[8]We use information from our human-annotated completion data as well as native speaker knowledge to construct an exhaustive list of valid completions per condition for this purpose.

3.4 Model Evaluation

All the models are evaluated by comparing the model output with the sentence completion data obtained from the native speakers; specifically, model output is evaluated in terms of the nature of the predicted verb class. We let $\mathbb{VC}$ denote the set of all verb-classes, $h(x)$ denotes the probability distribution of verb-class predictions made by humans, and $m(x)$ denotes the corresponding distribution of the model. We measure KL-divergence between these two distributions, replacing zero probabilities with a fixed value[9] ($= 10^{-5}$); this is shown in (1)

$$KLp(h||m) = KL(h||m') \tag{1}$$

where KL denotes the KL-divergence and m' is a distribution such that $m'(x) = \max(m(x), 10^{-5})$ for each $x \in \mathbb{VC}$.

Apart from this primary measure, we use two other metrics F and D to quantify the span and quality of model predictions with respect to the predicted verb classes, respectively, in order to better understand these characteristics of each model (see Section 6.1).

Further, to ascertain a qualitative understanding of the model performance, we also evaluate each model on the basis of the following characteristics that are displayed in the completion data discussed in Section 2:

- Deterioration in the number of grammatical completions on the 3-NP conditions compared to the 2-NP conditions

[9]It is equal to the minimum probability that we allowed in our model predictions

- Within the grammatical completions, a preference for simpler structures as opposed to complex or embedded constructions
- Exhibition of similar types of errors as humans; for example, in 3-NP conditions, N1-N2 errors, as well as a sensitivity to subject primacy with the Ergative case.

For the 3-NP conditions, we classify errors into types based on their compatibility with a 2-NP sub-context (N1-N2, N1-N3, N2-N3). For example, an error type of N1-N2 would mean that the corresponding ungrammatical prediction is compatible only with first two NPs and not the full 3-NP context. This scheme follows the error types found in the completion data discussed in Section 2. Additionally, see Section 2 of the supplementary material for examples of various errors.

4 N-gram Based Surprisal Model

In order to evaluate the *adaptability hypothesis* where the prediction of upcoming verb is driven by local nominal arguments, we implement an n-gram language model using the data discussed in Section 3.2. Such models are typically used to compute the surprisal metric (Hale, 2001; Levy, 2008) given local context (e.g., Levy et al., 2012). Recall that we have at most 3 NPs as the preverbal context, and therefore, we use a 4-gram model so that the model has access to the complete context in a given condition to make a verbal prediction. Unlike the models discussed in Section 5, the preverbal context in this model is free of noise.

5 Lossy-context Surprisal Models

In this section, we discuss two models to test the *noisy channel hypothesis*. As stated in Section 1, the underlying assumption is that human communication is noisy (Gibson et al., 2013; Kurumada and Jaeger, 2015) and the comprehender has to reinterpret the input to make prediction about upcoming linguistic material. In order to evaluate this hypothesis, we implement different versions of the *lossy-context surprisal* metric (Futrell et al., 2020). Lossy-context surprisal holds that processing difficulty at a word in a context is proportional to the surprisal of a word given a *lossy memory representation* of the context. The two models discussed in sections 5.1 and 5.2 differ in their noise functions that affect the interpretation of the preverbal context.

For the current investigation, lossy-context surprisal is extended to model the sentence-completion task. The word with the highest probability in a given context is assumed to be most likely to complete the sentence (cf. Staub et al., 2015; Levy, 2008; Smith and Levy, 2013).

As noted by Futrell et al. (2020), the lossy surprisal model is not representation-agnostic. Its predictions are dependent on a noise distribution (M). One can then obtain:

$$p(w|r) \propto \sum_{c} p_M(r|c)p(c)p_L(w|c), \qquad (2)$$

where w is the predicted word and r is the result of adding noise to the context c. Here, we consider L to be a 4-gram model, same as the one discussed in Section 4. Moreover, for $c = w_1 w_2 \cdots w_n$ we calculate $p(c)$ also using L

$$p(c) = \prod_{i=1}^{n} p_L(w_i|w_{i-3}w_{i-2}w_{i-1})$$

In addition, if $|c| = n \leq 2$, we don't add any noise to the context and simply use the n-gram model L for prediction. In other words, if $c = w_1 w_2$ or $c = w_1$, then we consider $p(w|r) = p_L(w|c)$. Since we only consider erasure-based noise distributions, this is done to ensure that the whole context is not lost during prediction. In order to get an average behavior of the model, we run the model 10 times and then take the top 50 predictions based on the total probability of each prediction. In other words, suppose a phrase s is predicted to follow a given preverbal arguments in a condition. Then, the total probability of s to be predicted in the given condition by the average model is equal to $\frac{1}{10}\sum_{i=1}^{10} p_i(s)$, where $p_i(s)$ denotes the probability of prediction s in the ith run. Note that if s is not predicted in the ith run, then $p_i(s) = 0$. In the next subsections, we present two models with different noise distribution.

5.1 Predictability Bias Noise (LC-Surp Pred-Bias)

We first consider a noise distribution such that the context is reconstructed based on the predictability of a sub-context. This is driven by the idea that reconstruction of context given a noisy input will be influenced by prior linguistic exposure (Futrell et al., 2020). When the input is less frequent, its reconstruction will be influenced by frequent linguistic patterns in the language. Note, however, that a single word is obviously more frequent than

two. Hence, we needed to control for the reduction in the size of the context that may arise due to this predictability bias. We do this by selecting sub-contexts based on their size with a preference to a larger size. Starting from the complete context, we thus iteratively reduce the size by 1 with a high probability ($d = 0.8$).[10] Thus, a sub-context of size m is considered with a probability d^{n-m} where m is the size of the corresponding context. Hence,

$$p_M(r|c) \propto d^{n-m} p_L(r) \tag{3}$$

5.2 Predictability Recency Noise (LC-Surp Pred-Rec)

We next consider a noise distribution which exploits both predictability bias as well as recency. It is well attested that recent input is easier to retrieve from memory compared to non-recent input (e.g., Lewis and Vasishth, 2005). The function therefore is motivated by the fact that while previous linguistic exposure should influence context reconstruction (Futrell et al., 2020), this reconstruction should bias recent linguistic material. In a way, this model combines the properties of the Predictability bias noise model and the n-gram surprisal model.

The conditional probability $p(r|c)$, here, thus can be seen as the multiplication of two parts - (a) predictability of r, $p_L(r)$; and (b) decaying erasure factor, $p_{rec}(r|c)$. Let $c = w_1 w_2 \cdots w_n$, $r = w_{i_1} w_{i_2} \cdots w_{i_k}$ for some n, k, then

$$p_M(r|c) \propto \prod_{j=1}^{n-k} f^{n-i_j} p_L(r), \tag{4}$$

where f is a constant fixed at 0.8.[11]

Thus, a context which is both predictable and can be formed from a recent subcontext is favored. The further a word is from the last uttered word, the lesser its likelihood of being a part of the reduced context r.

6 Results

Table 4 compares the verb class results for the three models discussed above. The key finding is that the values of KLp for the LC-Surp Pred-Rec model is lower than the other models for most of the conditions. This suggests that the model performs better in capturing the verb class distribution found in the human data.

[10] We also evaluated the model with $d = 0.9$ but the model with $d = 0.8$ gave better results.

[11] Following the value fixed for d in Section 5.1.

Condition	4-gram	LC-Surp Pred-Bias	LC-Surp Pred-Rec
ne-ko-se	6.05	5.97	**3.93**
ne-se-ko	7.00	9.14	**5.32**
ko-ne-se	9.40	**9.39**	9.40
ko-se-ne	8.25	8.53	**8.24**
se-ko-ne	5.38	7.87	**5.35**
se-ne-ko	8.57	8.52	**8.37**
Average	7.44	8.24	**6.77**

Table 4: Comparison of the considered models for each condition based on the KLp metric (Equation 1) rounded to 2 places. Smaller (bold) means better.

In order to test if the improvement seen in the LC-Surp Pred-Rec model is indeed significant, we also performed the chi-square test to see if the categories of verb class predicted in the LC-Surp Pred-Rec model were significantly different from other models. Results showed that this was indeed true – categories of verb classes in the LC-Surp Pred-Rec model were significantly different ($p < 0.05$) from both 4-gram model and the LC-Surp Pred-bias model.[12]

KLp provides a measure to quantify the divergence between the human and model prediction distributions. However, the nature of this divergence is still unclear. In order to understand the output of the models better, we evaluate them on some additional metrics. Finally, we report a qualitative analysis of the model output.

6.1 Span and Quality of the Models

In this section we assess the span and quality of the predictions made by the models when compared to the human data.

The `span` of verb prediction made by the model can be computed by the proportion of human distribution that the model misses on. Formally,

$$F(h||m) \propto \sum_{\substack{x \in \mathbb{VC} \\ m(x)=0}} h(x) \tag{5}$$

Since the model will not be able to predict all verb classes that humans produce, we formulate a metric to evaluate the `quality` of the predictions that the model makes. For this, we restrict the verb classes to only those that are predicted by the model and find the KL-divergence (Kullback and Leibler,

[12] See Section 5 of the Supplementary material for details.

Condition	4-gram		LC-Surp Pred-Bias		LC-Surp Pred-Rec	
	F	D	F	D	F	D
ne-ko-se	0.58	2.42	0.58	2.30	**0.31**	**2.15**
ne-se-ko	0.68	2.12	0.94	**0.50**	**0.31**	4.06
ko-ne-se	0.96	0.33	0.96	0.35	0.96	**0.25**
ko-se-ne	**0.87**	0.32	0.90	0.37	**0.87**	**0.23**
se-ko-ne	0.51	2.05	0.83	**0.96**	**0.43**	2.87
se-ne-ko	0.90	0.23	0.90	0.35	**0.88**	**0.15**
Average	0.75	1.99	0.85	**1.63**	**0.63**	2.91

Table 5: Comparison of the considered models for each condition based on the metrics F, D as defined in Equations 5, 6. Smaller means better (bold represents the best in that row for each metric).

1951) on those verb classes between the model and the human; this is shown in (6)

$$D(h||m) = \sum_{\substack{x \in \mathbb{VC} \\ m(x) \neq 0}} h'(x) \log \frac{h'(x)}{m(x)} \quad (6)$$

where $h'(x)$ is normalized from $h(x)$ after removing x where $m(x) = 0$.

Note that higher the F, lower is the model's span; and similarly, higher the D, lower is its quality of predictions (as compared to humans). Table 5 shows that for both F and D, the LC-Surp Pred-Rec model consistently outperforms the LC-Surp Pred-Bias and the 4-gram surprisal model. This suggests that when compared to the human data, the LC-Surp Pred-Rec is better in predicting the valid verb class both in terms of span and the quality of the predictions.

6.2 Qualitative Analysis

In order to interpret the metrics mentioned in Table 5, we did a detailed analysis of the model output in terms of the nature of verb class and the type of prediction errors. This is summarized in Table 6. One can note that

- Grammaticality in all models drops in 3-NP conditions as compared to 2-NP conditions, in line with the human data (cf. Section 2) [13].
- The models prefer simple outcomes, and largely predict DT, $CAUS$ (grammatical) and T, N DT (ungrammatical).

Investigating the reason for the better span of the Pred-Rec model, we find that it is primarily due to the important *T DT* verb class. This embedded structure is often used by humans, and neither of the 4-gram or the Pred-Bias model managed to predict it; thus, we can link the better span numbers of the Pred-Rec model to an observable improvement in the nature of verbal predictions.

We also study the error types made by the models and compare them to human errors. The 4-gram model by its nature is only capable of making the locally coherent N2-N3 errors, whereas both the Pred-Bias and Pred-Rec models produce N1-N3 and N1-N2 errors as well. However, while the human data was sensitive to the subject primacy effect – presence of Ergative case-marker never lead to passive verb completion; none of the models is able to fully replicate this pattern. However, the 4-gram model produces the least percentage of passives, followed by the Pred-Rec model. See Section 6 of the supplementary material for more details about error types.

7 Discussion

Results show that the Lossy context surprisal model with Predictability Recency Bias noise performs best in terms of the distribution of predicted verbs and the error types vis-à-vis the completion data. This provides support for the *noisy channel hypothesis* and poses a challenge to the *adaptability hypothesis*. In addition, the comparison of the two lossy surprisal models sheds light on the nature of the noise during the reconstruction process.

Results show that qualitatively all the models capture the completion data to a certain extent (see, Section 6.2). At the same time, overall the noisy context models performed better than the n-gram model in two clear ways. First, the models were able to capture the differential nature of case-marker combination in a limited context. This leads to better coverage of error sources (both in terms of errors made and not made). Second, the models were therefore also better at making better verb predictions compared to the n-gram model. In particular, the overall success of the Pred-Rec model showed that reconstruction of the noisy context in influenced by both past exposure of preverbal sub-context and the recency of the context (cf. Futrell et al., 2020). Put differently, the reconstruction of the context is driven by sub-strings that are more frequent (e.g., ne-ko) and that are closer to the verb. Critically, this shows that the reconstruction process is not random.[14]

[13] See Section 5 of the supplement for actual percentages.

[14] In addition to the two noise functions reported in Sec-

Characteristic	4-gram	LC-Surp Pred Bias	LC-Surp Pred-Rec
Gm% (2-NP) > Gm% (3-NP)	**Yes**	**Yes**	**Yes**
Grammatical classes	*DT*, *CAUS*	*DT*, *CAUS*	***DT*, *CAUS*, *T DT***
Embeddings predicted	No	No	**Yes**
% of passives	**2.5%**	4.2%	3.1%
Errors made	Only N2 N3 errors	**All error types**	**All error types**

Table 6: Qualitative analysis of the models' predictions. The best/desired outcomes appear in bold font. Gm% denotes the proportion of grammatical completions predicted. High % of passives signifies insensitivity to subject primacy.

While the performance of the predictability recency model is good, it suffers from three issues (a) it overestimates the number of errors made by humans, (b) its overall coverage for various verb class is low, and (c) it is insensitive to subject primacy. The model is able to successfully predict verb phrase involving no clausal embedding, and to a limited extent, those with embeddings. While certain complex structures such as *N DT DT*, predicted rarely by humans, are dropped entirely by the model, its prediction for the *T DT* structure which is frequent in the completion data is not that high. An investigation into the data also shows a scarcity of training examples that exhibit an animate 3-NP context followed by such *T DT* continuations.[15] One reason for this could be the size of the training data, currently 5 million sentences; future work can train on a larger data set. Another possibility is that certain patterns in the human data are not captured in the written corpus used for training and requires a dialogue corpus. Unfortunately, such a corpus currently does not exist for Hindi and attempts to modeling using such a data will have to wait its availability. Relatedly, Staub et al. (2015) argue that prediction based on corpus frequency of syntactic information may not be able to fully capture the notion of preactivation during the completion task. Hence, future work will need to incorporate other sources of information. Finally, the results show that the 4-gram model is more sensitive to subject primacy. This is because, the 4-gram model (unlike noisy context models) has access to the N1 features when making predictions. It can thus correctly use the N1 case feature to avoid predicting passive verbs. This suggests that a noise function relying only on local information will be limited in accounting for the current data.

tion 5, we also investigated a purely random noise function. Due to space constraint, details of this model have been mentioned as supplementary material (Section 7).

[15]See Section 3 of the supplementary material for more details on training data.

The current work provided the first set of detailed results towards modeling clause final verb prediction in an SOV language. The work demonstrated the effectiveness of lossy surprisal models and probed the nature of the noise function during the reconstruction process. In addition to the quantitative analyses demonstrating the success of the Predictability Recency lossy surprisal model, a key contribution of the work was that it highlighted the nature of model's closeness to the human data, both in terms of verb class prediction and the error type. Overall, the results support the proposals that highlight the detrimental effect of increased complexity of the preverbal linguistic material in SOV languages (e.g., Gibson et al., 2013; Ueno and Polinsky, 2009; Ros et al., 2015; Yadav et al., 2020). Future models need to explore other noise functions to investigate the interaction of context predictability with recency as well as primacy of non-local information (e.g., subject). Further, these models need to be tested to investigate the effect of distance (e.g., Vasishth and Lewis, 2006) and structural complexity (Vasishth et al., 2010) on verbal prediction in SOV languages.

8 Conclusion

We implemented three models to predict clause final verbs in Hindi. Model outputs were compared with verb predictions of native speakers of Hindi using quantitative measures as well as qualitatively. Results show that the model that uses limited preverbal context with a predictability recency bias noise function captures the distribution of human data best. The success of this model is consistent with the idea that the reconstruction of the noisy context during prediction is influenced by prior linguistic exposure and that this process interacts with recency of input. These results support the *noisy channel hypothesis* to language comprehension.

References

Arpit Agrawal, Sumeet Agarwal, and Samar Husain. 2017. Role of expectation and working memory constraints in hindi comprehension: An eyetracking corpus analysis. *Journal of Eye Movement Research*, 10(2):1–15.

G. T. Altmann and Y. Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–256.

Apurva and Samar Husain. 2020. Parsing errors in hindi: Investigating limits to verbal prediction in an sov language. *In submission*.

Riyaz Bhat. 2017. *Exploiting Linguistic Knowledge to Address Representation and Sparsity Issues in Dependency Parsing of Indian Languages.* Ph.D. thesis, IIIT Hyderabad India.

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 (LAW III)*, pages 186–189.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marisa Ferrara Boston, John T Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

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

F. Ferreira and N. D. Patson. 2007. The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1:71–83.

Angela D Friederici and Stefan Frisch. 2000. Verb argument structure processing: The role of verb- specific and argument-specific information. *Journal of Memory and Language*, 43(3):476–507.

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.

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088.

Alvin Grissom II, Naho Orita, and Jordan Boyd-Graber. 2016. Incremental prediction of sentence-final verbs: Humans versus machines. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 95–104, Berlin, Germany. Association for Computational Linguistics.

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*, pages 1–8. Association for Computational Linguistics.

Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. 2014. Strong Expectations Cancel Locality Effects: Evidence From Hindi. *PloS one*, 9(7):e100986.

Jana Häussler and Markus Bader. 2015. An interference account of the missing-vp effect. *Frontiers in Psychology*, 6:766.

Itisree Jena, Riyaz Ahmad Bhat, Sambhav Jain, and Dipti Misra Sharma. 2013. Animacy annotation in the Hindi treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 159–167, Sofia, Bulgaria. Association for Computational Linguistics.

A. J. Knoedler, K. A. Hellwig, and I. Neath. 1999. The shift from recency to primacy with increasing delay. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(2):474–487.

Lars Konieczny. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Gina R. Kuperberg and T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1):32–59.

Chigusa Kurumada and T. Florian Jaeger. 2015. Communicative efficiency in language production: Optional case-marking in japanese. *Journal of Memory and Language*, 83:152 – 178.

M. Kutas and S.A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307:161 – 163.

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

Roger Levy, Evelina Fedorenko, Mara Breen, and Edward Gibson. 2012. The processing of extraposed structures in english. *Cognition*, 122(1):12–36.

Roger Levy and Frank Keller. 2013. Expectation and locality effects in german verb-final structures. *Journal of memory and language*, 68(2):199–222.

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3):375–419.

Steven G. Luke and Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22 – 60.

W. Marslen-Wilson. 1973. Linguistic structure and speech shadowing at very short latencies. *Nature*, 244:522–523.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Idoia Ros, Mikel Santesteban, Kumiko Fukumura, and Itziar Laka. 2015. Aiming at shorter dependencies: the role of agreement morphology. *Language, Cognition and Neuroscience*, 30(9):1156–1174.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka, Japan. The COLING 2016 Organizing Committee.

N. J. Smith and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

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

A. Staub and Jr. Clifton, C. 2006. Syntactic prediction in language comprehension: Evidence from either ... or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32:425–436.

A. Staub, M. Grant, L. Astheimer, and A. Cohen. 2015. The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82:1–17.

Whitney Tabor, Bruno Galantucci, and Daniel Richardson. 2004. Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355 – 370.

W. Taylor. 1953. 'cloze' procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

J. Trueswell, M. K. Tanenhaus, and S. M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318.

Miekko Ueno and Maria Polinsky. 2009. Does headedness affect processing? a new look at the vo–ov contrast. *Journal of Linguistics*, 45(3):675–710.

Shravan Vasishth and Richard L Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, pages 767–794.

Shravan Vasishth, Katja Suckow, Richard L. Lewis, and Sabine Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: a cross-linguistic corpus study. *Cognitive Science*, 44(4).

Hiroko Yamashita. 1997. The effects of word-order and case marking information on the processing of japanese. *Journal of Psycholinguistic Research*, 26(2):163–188.