Supplementary Material
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

1 Quick Overview of Hindi

Hindi is a Subject-Object-Verb (SOV) order language. It allows for flexibility in word order, owing to the fact that nominal case-markers can signify syntactic roles. In addition, the verb agrees with its highest unmarked argument; if all arguments are case-marked then the verb takes default agreement features. For example, in (1) below, the subject (*Pooja*) and indirect object (*Urmila*) are case-marked with the Ergative and Ablative case-markers *-ne* and -se respectively. The verb *lii* 'took' agrees with the unmarked object *kitaab* 'book'.

(1) pooja-**ne** urmila-**se** kitaab lii
Pooja-ERG Urmila-ACC book took
Pooja took a book from Urmila

Due to the permissible free word order, the preverbal subject, indirect object and direct object can occur in various orders. See, Kachru (2006), for a detailed description on word order, case, agreement and other syntactic properties of the language.

2 Further Details on the Completion Study

Examples (2) show the conditions where 2 preverbal nouns preceded the target verb; similarly, Examples (3) show the conditions where 3 preverbal nouns preceded the target verb. In the examples, *ne* is the Ergative case-marker, *ko* is the Accusative case-marker and *se* is the Ablative case-marker. 36 native speakers participated in the 3-NP condition experiments. 25 native speakers participated in the 1-NP and 2-NP condition experiments.

(2) a. ne-ko
pooja-**ne** urmila-**ko** ...
Pooja-ERG Urmila-ACC ...

b. ne-se
pooja-**ne** urmila-**se** ...
Pooja-ERG Urmila-ABL ...

c. ko-ne
pooja-**ko** urmila-**ne** ...
Pooja-ACC Urmila-ERG ...

d. ko-se
pooja-**ko** urmila-**se** ...
Pooja-ACC Urmila-ABL ...

e. se-ko
pooja-**se** urmila-**ko** ...
Pooja-ABL Urmila-ACC ...

f. se-ne
pooja-**se** urmila-**ne** ...
Pooja-ABL Urmila-ERG ...

(3) 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 ...

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

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

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

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

Using a rating study, Apurva and Husain (2020) also find that many errors (e.g., the locally coherent N2 N3 error) can be deemed grammatical illusions, i.e., such parses were rated similar to their grammatical counterpart (cf. Gibson and Thomas, 1999). In addition, the errors show a subject primacy effect (Häussler and Bader, 2015) where the case-marker of the subject is not forgotten while making a prediction. This is particularly true for the Ergative (-ne) case which requires a perfective aspect on the verb. The completion data shows that in such cases a verb with a passive morphology is never predicted.

3 Raw Data: Verb Phrase Heuristics

We analysed the raw data to test its suitability as input for training 4-gram models, for the task of

verb prediction. We were chiefly concerned with gauging the distance between the required condition environments and the verb phrase, since the ability of the language model to learn valid verbal completions would decrease with the intervening distance. Similarly, it would also be affected by prepositional or adverbial interveners within the verb phrase itself. Table 1 shows a summary of the raw data. 58.33% of the conditions show an average verb phrase distance larger than 4; this, coupled with the average internal intervener distance, indicated that this data would not be very tractable to a 4-gram model.

The numbers in the table were calculated by extracting sentences with the given conditions from the corpus, obtaining a dependency parse, and identifying head verb, dependent arguments and adjuncts, etc. The verb phrase was identified using a set of valid POS tags and dependency relations, in a way that preserved complex predicates. Prepositional, adjectival, and adverbial phrases were classed as interveners.

4 Sentence Simplification

The simplification pipeline was motivated by the above analysis of the raw data. Table 2 shows the equivalent numbers on the simplified data. The average verb phrase distance has dropped to less than 1 for 83.33% of conditions, indicating that a 4-gram model has a good chance of capturing relationships between the verb and its argument context. Similarly, internal intervener distance, or the average number of intervening words within the verb phrase itself, has also dropped to near zero numbers for all conditions.[1]

We use manual annotation to slot the extracted verb phrases into classes, in combination with string matching. Table 3 shows the results of this manual classification.[2]

4.1 Testing of the Simplification Pipeline

We want to ensure two things the post-simplification step:

- Grammaticality is preserved
- Verb classes are preserved

In order to test this, we randomly choose 25 sentences from the raw data and pass them through the pipeline. We manually verify the grammaticality and quality of the resulting simplified sentences, also ensuring that the verb class is preserved from the original.

Example of simplification:

Aakhir kaaphi der-ke pratiksha
Finally, enough time-ACC-GEN wait
karne-ke baad pareshani-ki sthiti-mein
do-INF-GEN after, trouble-GEN situation-in
A-**ne** A-**ko** phone-par **khabar di**
A-NOM A-ACC phone-on **news give-PT**

↓

A-**ne** A-**ko** **khabar di**
A-ERG A-ACC **news give-PT**

Investigating cases where grammaticality/verb phrase is disrupted post-simplification, we find that it can be caused due to automatic parsing errors, and its effect on the the simplification algorithm. For example, the simplification algorithm may sometimes miss out phrasal verbs such as the on the following example due to the lack of relevant information from the automatic parser, where a prepositional phrase is essential to the meaning of the verb:[3]

A-**ne** A-**ko** **hirasat-mein liya-hai**
A-ERG A-ACC **custody-PSP take-AUX-P.Pf**

↓

A-**ne** A-**ko** **liya-hai**
A-ERG A-ACC **take-AUX-P.Pf**

Next, we need to verify that verb classes of interest are preserved through simplification, so that the language models have exposure to them. We do this by randomly choosing two sentences per verb class from the raw data, passing it through the simplification pipeline and manually verifying that the verb class is maintained.

Examples:
$N\ DT$

Ahmedabad: A-**ne** A-**ko** doosre
Ahmedabad: A-ERG A-ERG second-ABL
twenty-twenty mukaable-mein 11
twenty-twenty competition-LOC-in 11
runon-se haraaya
run-PL-ABL defeat-PT

[1] Note that the numbers for the total per condition have also changed. The extraction of these conditions is based on simple string matching rather than more robust dependency relations. Therefore, it is possible that we earlier identified a sequence as a condition despite some of its parts being adjunct rather argument material. The simplification, by removing these, would then disrupt the condition. We test the simplification to ensure that it does not disrupt argument condition sequences.

[2] Incoherent or non-verb utterances were labelled as $Other$, e.g., 'N N'

[3] This happens when the parser does not identify a phrasal verb as such with the attendant dependency relations.

Condition	Type	Total	VP Len < 3	VPD < 2	VPD < 4	Internal Intervener > 1	Avg VP Length	Avg VP Dist	Avg Int Intervener
ne se	2 NP	806	775	338	232	26	1.59	3.24	0.2
ko se	2 NP	98	93	28	19	12	1.18	4.22	0.67
ne ko	2 NP	1381	1281	282	455	76	1.34	4.52	0.31
se ne	2 NP	245	209	46	32	30	1.31	7.46	0.82
ko ne	2 NP	373	346	51	150	18	1.38	4.06	0.29
se ko	2 NP	67	67	8	22	0	1.36	5.15	0.03
ko se ne	3 NP	0	0	0	0	0	0	0	0
ne ko se	3 NP	21	21	2	14	0	1.1	2.57	0
se ne ko	3 NP	8	2	0	2	6	0.25	3.75	6
ne se ko	3 NP	10	10	8	0	0	1.6	1.2	0
ko ne se	3 NP	2	2	0	0	0	2	4	0
se ko ne	3 NP	6	6	0	0	0	2	9	0

Table 1: Verb Phrase heuristics: VPD = Verb phrase distance from the condition, measured in number of intervening words, Internal Intervener = Number of adjunct or non-arguments words within the span of the VP, measured in number of words. ne=Ergative, ko=Accusative, se=Ablative.

↓

A-**ne** A-**ko** **runon-se haraaya**
A-ERG A-ERG **run-PL-ABL defeat-PT**

$N\ CAUS$

A-**ko** A-**ne** court mein pesh karne
A-ACC A-ERG court in present-INF -
ke baad **jel bhijva diya**
PSP after **jail send-CAUS AUX**

↓

A-**ko** A-**ne** **jel bhijva diya**
A-ACC A-ERG **jail send-CAUS AUX**

This also confirms that the statistics on verb classes in the simplified data in Table 3 are representative of the verb classes in the raw data. They can therefore be used to contextualize the predictions of our models. The low numbers of 3-NP conditions mean that our models have very little exposure to these to begin with. Note the almost complete absence of TDT type VPs, which are on the contrary among the most frequent completions made by humans in response to 3-NP conditions. On the other hand, we have relatively good numbers for T type VPs following 2-NP conditions, and a distribution over DT and $CAUS$ type VPs for 3-NP conditions. This is reflected in the results for our models, although varying, of course, based on the nature and underlying hypothesis of the specific model.

5 Results: Significance Testing

The models share some characteristics in common, e.g. the primary verb classes produced, certain error types, and deterioration of grammaticality from 2-NP to 3-NP conditions.[4]

In order to make a meaningful claim based on the better performance of the LC-Surp Pred-Rec as compared to the other two models, we recognize the need to perform significance testing on the KLp metric values. Since any direct test of this sort is infeasible for obvious reasons, we demonstrate, instead, that the distribution over verb classes produced by each model per condition is significantly different from that produced by the others. In essence, we show that the premise behind each model creates output behaviour that is significantly different from others, where output behaviour is reasonably encoded as the probability distribution over verb classes given a condition. If this is true, then the lower Klp value of the distribution as a whole against the human distribution is indeed an indicator of superiority of the model and its associated premise.

We annotated the model lexical completions into verb classes by manual annotation, thus obtaining a probability distribution over verb classes given a model given a condition. We sample each model output distribution per condition 500 times[5] in order to obtain the sample set of a categorical variable, given condition. Then, we perform the $chi-2$ significance test between the obtained numbers for each pair of models, given each condition.[6] We obtain p-values below the 0.05 threshold for all conditions for all three models pairs except one condition-model triplet as shown in Table 4.

[4]We have percentage of grammatical completions for 4-gram: 54.1%, 41.1.%, Pred-Bias: 54.1%, 44.9%, and Pred-Rec: 49.4%, 39.6% for 2-NP and 3-NP conditions respectively.

[5]This number is chosen since it is roughly the number of human data points that we have. We get similar results for lower numbers as well, although the test is increasingly unreliable at low numbers according to some standard heuristics.

[6]We simply smoothed zero-values with 1

Condition	Type	Total	VP Len < 3	VPD < 2	VPD < 4	Internal Intervener > 1	Avg VP Length	Avg VP Dist	Avg Int Intervener
ne se	2 NP	430	428	397	31	0	1.6	0.32	0.01
ko se	2 NP	48	48	43	4	0	1.27	0.96	0
ne ko	2 NP	1275	1263	896	346	2	1.45	0.97	0.01
se ne	2 NP	15	15	0	11	0	1.4	3	0
ko ne	2 NP	336	336	260	70	0	1.44	0.76	0
se ko	2 NP	7	7	4	3	0	1.86	1.14	0
ko se ne	3 NP	0	0	0	0	0	0	0	0
ne ko se	3 NP	32	32 30	2	0	1.34	0.25	0	
se ne ko	3 NP	2	2	2	0	0	1	1	0
ne se ko	3 NP	7	7	3	4	0	1.86	1.71	0
ko ne se	3 NP	14	14	12	2	0	2	0.29	0
se ko ne	3 NP	0	0	0	0	0	0	0	0

Table 2: Simplified Data. VPD = Verb phrase distance from the condition, measured in number of intervening words, Internal Intervener = Number of adjunct or non-arguments words within the span of the VP, measured in number of words. ne=Ergative, ko=Accusative, se=Ablative.

Condition	Type	Total	CAUS	DT	IN	N+CAUS	N+DT	N+Pass	N+T	N+T+DT	Pass	T	T+DT	Other
ne se	2 NP	430	0	35	2	0	77	0	1	4	0	139	2	170
ko se	2 NP	48	0	2	0	0	6	2	2	0	16	18	0	2
ne ko	2 NP	1275	40	252	0	0	322	0	7	2	58	527	14	53
se ne	2 NP	15	0	1	0	0	1	0	0	0	0	11	0	2
ko ne	2 NP	336	3	67	0	0	43	0	0	0	28	186	3	6
se ko	2 NP	7	0	1	0	0	0	0	0	0	0	2	0	4
ko se ne	3 NP	0	0	0	0	0	0	0	0	0	0	0	0	0
ne ko se	3 NP	32	0	5	0	0	0	0	9	2	0	16	0	0
se ne ko	3 NP	2	0	2	0	0	0	0	0	0	0	0	0	0
ne se ko	3 NP	7	2	0	0	0	4	0	0	0	0	1	0	0
ko ne se	3 NP	14	0	0	0	2	2	0	0	0	0	6	0	4
se ko ne	3 NP	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3: Verb class statistics in the simplified data. *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, Pass=Passive. ne=Ergative, ko=Accusative, se=Ablative.

Condition	4-gram , Pred-Rec	4-gram , Pred-Bias	Pred-Bias, Pred-Rec
ne ko se	1.02e-10	1.42e-12	3.34e-33
ne se ko	6.69e-16	6.90e-12	8.88e-10
ko ne se	0.0111	**0.1388**	7.88e-05
ko se ne	0.0001	2.37e-24	1.87e-27
se ne ko	0.0049	1.05e-13	1.16e-15
se ko ne	9.00e-34	2.60e-16	1.87e-15

Table 4: p-values for chi-2 significance testing. Insignificant values, according to a threshold of 0.05 appear in bold font.

6 Results: Error Types across Various Models

Figures 1, 2 and 3 compare the error types found in the human data and in various models. As shown in Figure 3, the lossy-context surprisal model with Predictability-Recency noise function performs the best in terms of capturing the nature of error types – compared to the n-gram model it predicts not just the N2-N3 errors but also N1-N2 and N1-N3 errors; compared to the LC-Surp Pred-Bias model it is able to predict the **N2-N3** errors for the **ne-se-ko** condition and the **N1-N3** errors for **se-ne-ko** condition.

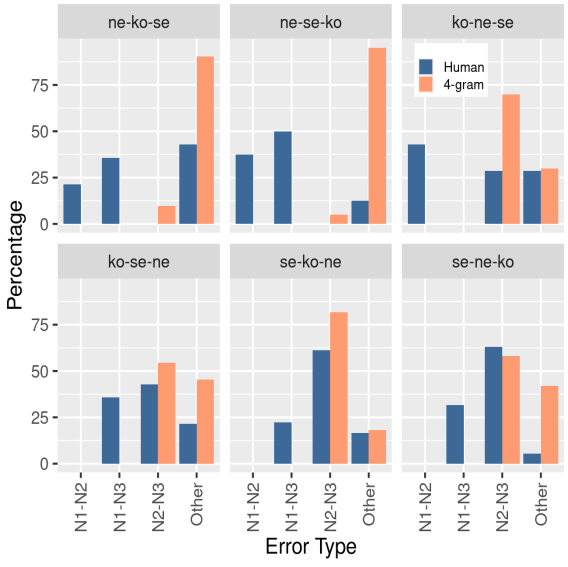

Figure 1: Comparison of the percentage of different error types between the 4-gram model and the humans for each condition.

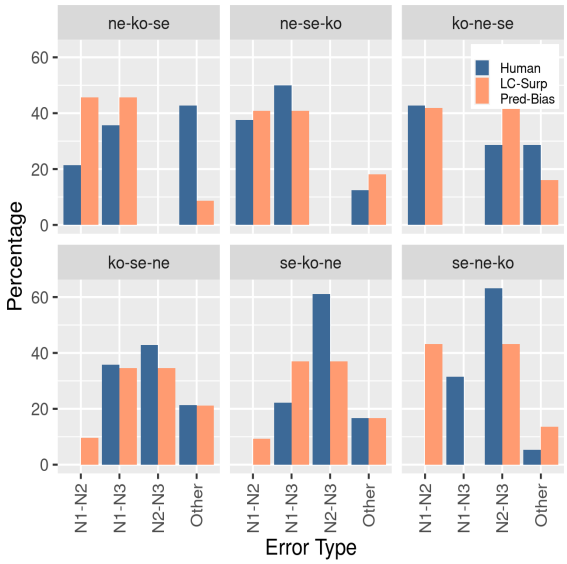

Figure 2: Comparison of the percentage of different error types between the Lossy-context Surprisal with Predictability bias noise and the humans for each condition.

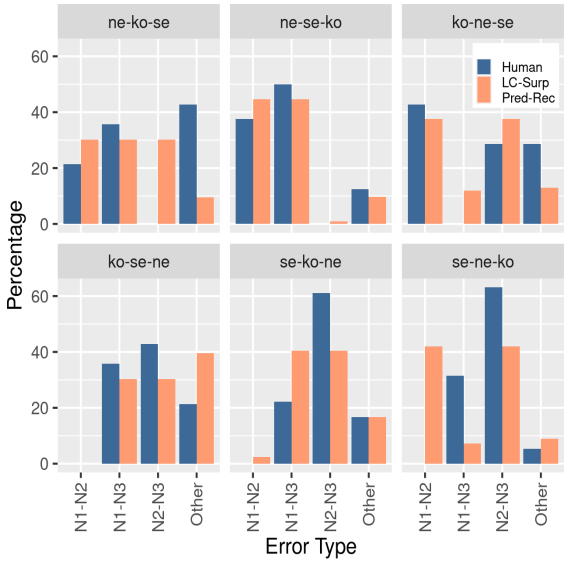

Figure 3: Comparison of the percentage of different error types between the Lossy-context Surprisal with Predictability-Recency noise and the humans for each condition.

7 Random Erasure Noise (LC-Surp Rand-Eras)

In this model, we consider a noise distribution such that the context is reduced by randomly deleting words from it. Each word in the context is deleted randomly with a constant probability e, which is fixed to be equal to 0.1 (Futrell et al., 2020).

$$p_M(r|c) \propto e^{|c|-|r|}(1-e)^{|r|} \quad (1)$$

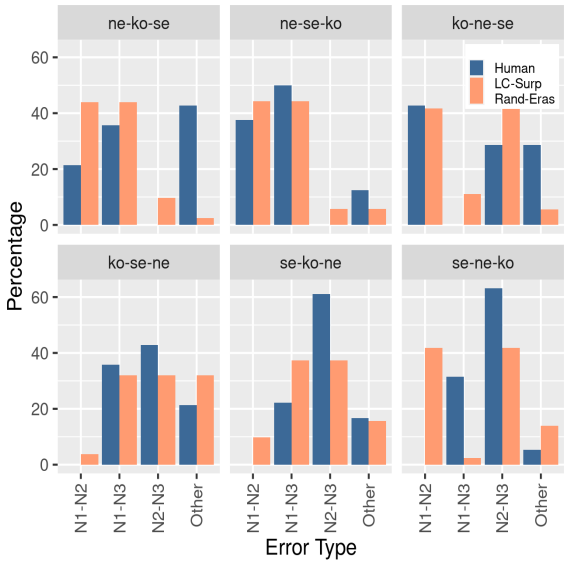

Figure 4: Comparison of the percentage of different error types between the Lossy-context Surprisal model with Random Erasure noise and the humans for each condition.

The results for the model show that 52% of all predictions were grammatical across all the 2-NP conditions while only 44.8% of the predictions were grammatical across the 3-NP conditions.

Similar to what was observed in the n-gram surprisal model, here too, the most frequent completions for the 3-NP conditions were simple (*DT*, *CAUS*, *T* etc.), in line with the completion study. We note that this model also fails to predict any kind of clausal embedding.

Figure 4 shows the classification of errors made by the model for each condition and how it compares with the completion errors. Now, the model succeeds in capturing various **N1-N2** and **N2-N3** errors made by humans across conditions. This happens at the cost of reducing the percentage of **N2-N3** errors in the **se-ko-ne** and **se-ne-ko** conditions.

References

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

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 and James Thomas. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammati-

cal. Language and Cognitive Processes, 14(3):225–248.

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

Y. Kachru. 2006. Hindi. John Benjamins Publishing Company, Philadelphia.