Cognition 223 (2022) 104959

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Locality and expectation effects in Hindi preverbal constituent ordering

Sidharth Ranjan [a,b], Rajakrishnan Rajkumar [b,*], Sumeet Agarwal [a,c]

[a] School of Information Technology, IIT Delhi, Hauz Khas, New Delhi 110016, India
[b] Department of Humanities and Social Sciences, IISER Bhopal, Bhauri, Madhya Pradesh 462066, India
[c] Department of Electrical Engineering, IIT Delhi, Hauz Khas, New Delhi 110016, India

ARTICLE INFO

Keywords:
Language production
Dependency length
Surprisal
Constituent ordering
Hindi

ABSTRACT

We investigate the relative impact of two influential theories of language comprehension, *viz.*, Dependency Locality Theory (Gibson, 2000; DLT) and Surprisal Theory (Hale, 2001; Levy, 2008), on preverbal constituent ordering in Hindi, a predominantly SOV language with flexible word order. Prior work in Hindi has shown that word order scrambling is influenced by information structure constraints in discourse. However, the impact of cognitively grounded factors on Hindi constituent ordering is relatively underexplored. We test the hypothesis that dependency length minimization is a significant predictor of syntactic choice, once information status and surprisal measures (estimated from *n*-gram *i.e.*, trigram and incremental dependency parsing models) have been added to a machine learning model. Towards this end, we setup a framework to generate meaning-equivalent grammatical variants of Hindi sentences by linearizing preverbal constituents of projective dependency trees in the Hindi-Urdu Treebank (HUTB) corpus of written text. Our results indicate that dependency length displays a weak effect in predicting reference sentences (amidst variants) over and above the aforementioned predictors. Overall, trigram surprisal outperforms dependency length and parser surprisal by a huge margin and our analyses indicate that maximizing lexical predictability is the primary driving force behind preverbal constituent ordering choices in Hindi. The success of trigram surprisal notwithstanding, dependency length minimization predicts non-canonical reference sentences having fronted direct objects over variants containing the canonical word order, cases where surprisal estimates fail due to their bias towards frequent structures and word sequences. Locality effects persist over the *Given-New* preference of subject-object ordering in Hindi. Accessibility and local statistical biases discussed in the sentence processing literature are plausible explanations for the success of trigram surprisal. Further, we conjecture that the presence of case markers is a strong factor potentially overriding the pressure for dependency length minimization in Hindi. Finally, we discuss the implications of our findings for the information locality hypothesis and theories of language production.

1. Introduction

A substantive body of work has put forth the view that language as a system evolved over time as a consequence of cognitive mechanisms and pressures related to language use and acquisition. According to this view, the structure of natural language is influenced by cognitive pressures related to production and comprehension (Hawkins, 2004), learnability (Christiansen & Chater, 2008) and communicative efficiency (Gibson et al., 2019; Jaeger & Tily, 2011). In this study, we investigate the relative impact of cognitive measures proposed by two influential theories of language comprehension, *viz.*, Dependency Locality Theory (Gibson, 2000; DLT) and Surprisal Theory (Hale, 2001; Levy, 2008) in predicting preverbal constituent ordering in written Hindi. Theories of real-time sentence comprehension include *expectation*-based and *memory*-based approaches depending on how cognitive resources are used (Levy, 2013). DLT focuses on *memory* – the utilization of resources for storing and retrieving linguistic representations. In contrast, Surprisal Theory focuses on *expectations* – the allocation of resources to alternate structures under conditions of uncertainty. Hawkins' pioneering work using corpus data from typologically diverse languages (Hawkins, 1994, 2000, 2004, 2014) demonstrated that languages tend to minimize dependency length (*i.e.*, the distance between syntactically related words). Subsequent large scale corpus studies based on dependency corpora of multiple languages belonging to distinct families have also confirmed the tendency of these languages to minimize dependency distance (Futrell et al., 2015; Liu, 2008).

* Corresponding author.
E-mail addresses: sidharth.ranjan03@gmail.com (S. Ranjan), rajak@iiserb.ac.in (R. Rajkumar), sumeet@iitd.ac.in (S. Agarwal).

https://doi.org/10.1016/j.cognition.2021.104959
Received 10 November 2020; Received in revised form 8 November 2021; Accepted 14 November 2021
Available online 25 January 2022

Hindi (Indo-Aryan language; Indo-European language family) has SOV canonical order along with a rich case-marking system and relatively free word order (Agnihotri, 2007; Kachru, 2006). The following examples from Mohanan and Mohanan (1994) illustrate word order flexibility in Hindi.[1]

(1) a. aaj maa=ne bacce=se kitaab padh-ne=ko kah-aa
today mother=ERG child=ACC book read-INF=ACC say-PFV.M.SG

Today the mother told the child to read the book.

b. aaj kitaab maa ne bacce se padhne ko kahaa
c. aaj bacce se kitaab maa ne padhne ko kahaa
d. bacce se kitaab maa ne padhne ko aaj kahaa
e. maa ne kitaab aaj bacce se padhne ko kahaa

In the above examples, permuting the preverbal constituents of the first sentence results in the remaining variant sentences, which express the same propositional content. As noted in pioneering work by Gambhir (1981), sentences having canonical order (like Example 1a above) can be considered to be neutral with respect to the preceding discourse context. Sentences with other orders (like the remaining sentences above), in contrast, are marked structures signifying various kinds of alternate emphases which may require context for complete interpretation. Subsequent work on scrambling in Hindi focused on information status factors related to discourse (Butt & King, 1996; Kidwai, 2000). The cited authors showed that scrambling is controlled by a host of factors related to discourse (topic, focus, background and completive information), semantics (definiteness and animacy) as well as prosody (Patil et al., 2008). Early work in Hindi sentence comprehension by Vasishth (2004) showed that the distance between the final verb and its preverbal arguments affected the processing difficulty of certain non-canonical word orders irrespective of the felicity of the preceding discourse context. Subsequently, Vasishth et al. (2012) reported that appropriate context can facilitate comprehension of certain types of non-canonical word orders. In this work, we factor in discourse considerations by adding an information status score to our statistical model.

The impact of cognitively motivated factors on syntactic choice in Hindi is thus underexplored, a theme we take up for further inquiry in this study. We test the hypothesis that dependency length minimization is a significant predictor of syntactic choice, once trigram surprisal, dependency parser surprisal and information status have been taken into account. In order to test the stated hypothesis, we incorporated predictors based on DLT integration costs and surprisal into a logistic regression model (Breslow & Clayton, 1993) aimed to distinguish Hindi reference sentences (like Example 1a above) and grammatical variants expressing the same idea, which were artificially generated (say Examples 1b–1e). In our study, we used sentences from the Hindi-Urdu Treebank (HUTB[2]) corpus of written Hindi (Bhatt et al., 2009). Thus we used sentences occurring in a given discourse context. Fig. 1 depicts the preponderance of short dependency lengths in the HUTB corpus. We estimated surprisal using word-based trigram language models as well as an incremental dependency parser (Agrawal et al., 2017). Information status was encoded in the form of a score quantifying previous mention in the previous sentence (*Given*) and first time mention (*New*) of both subjects and objects in sentences containing both these constituents. Our objective is to test theories of comprehension on syntactic choice in written data, which is a result of language production. Written text is often edited to facilitate comprehension for readers. Moreover, production choices in spontaneous speech are often guided by comprehension considerations as encoded in Levelt's (1989) model of language production, which has a self-monitoring component (perceptual loop connecting the production and comprehension systems). Speakers comprehend their own speech and make necessary adjustments prior to uttering (internal self-monitoring) as well as parse their own articulated utterances (external self-monitoring). For these reasons, we interpret our results by connecting them to established findings in sentence comprehension.

In this work we show that dependency length is a significant, but weak predictor of HUTB corpus sentences (over meaning-equivalent and grammatical variants) in a regression model containing information status and surprisal-based controls. Our experiments also revealed that trigram surprisal outperformed both dependency length and dependency parser surprisal by a huge margin in terms of regression model accuracy in predicting reference sentences over variants. In addition to broad coverage analyses involving the entire dataset, inspired by prior work (Husain et al., 2014; Vasishth, 2004), we also examined the impact of dependency length on two constructions, *viz.*, conjunct verbs and non-canonical word orders. Dependency length had a weak effect over and above surprisal in Hindi conjunct verbs. Locality effects emerged in a few cases where surprisal estimates performed poorly. Husain et al. (2014) showed a similar pattern for reading times in sentence comprehension for this construction. Further, dependency length minimization was effective in distinguishing between reference sentences containing direct object fronted structures (non-canonical order) and variants having the canonical order. Surprisal estimates performed poorly in detecting reference sentences with non-canonical orders due to their bias towards more frequent orders. Thus it is conceivable that memory considerations dominate in rare structures, along the lines of Hindi sentence comprehension work described earlier (Vasishth, 2004). Further studies based on the Potsdam-Allahabad corpus of Hindi reading times also attested the effect of DLT-based integration costs on various eye tracking measures, even in the presence of low-level controls like syllable length, unigram and bigram word frequencies (Agrawal et al., 2017; Husain et al., 2015). Going beyond the cited studies, we show that the effect of dependency length minimization went above and beyond the preference for *Given-New* orders encoded by the information status predictor. Future inquiries using spontaneous production experiments and speech corpora can investigate whether our results on written text hold for language production.

Our main contribution is that, using ecologically valid data, our study provides independent converging evidence about locality and expectation effects, consistent with previous proposals (described above), from a language other than English. We designed and executed a computational platform to test hypotheses related to word order using algorithms and software tools in vogue in Natural Language Generation. Using the above framework, we thus seek to expand the typological basis of influential theories of syntactic choice. In a comprehensive cross-linguistic survey of language production, Jaeger and Norcliffe (2009) showed that theories of language processing have been formulated by studying a small number of languages, an approach which incurs the risk of positing universal mechanisms of processing based on data from a few sources. They emphasize that data from diverse languages are essential for hypothesis testing as well as revising current theories of language processing. The structure of this paper is outlined below. Section 2 provides background information for the study. Section 3 gives a detailed description of our methods and data and Section 4 shows our results. Section 5 then discusses factors counteracting dependency length minimization in Hind and reasons for the overwhelming success of trigram surprisal. The remainder of the section also discusses the implications of our findings for information locality hypothesis and language production. Finally, Section 6 presents our conclusions.

[1] We follow the Leipzig Glossing Rules (https://www.eva.mpg.de/lingua/resources/glossing-rules.php/) in this paper: ACC: accusative; ERG: ergative; FUT: future; INF: infinitive; INS: instrumental; M: masculine; PFV: perfective; PL: plural; PROG: progressive; PRS: present; PST: past; SG: singular.

[2] http://verbs.colorado.edu/hindiurdu/

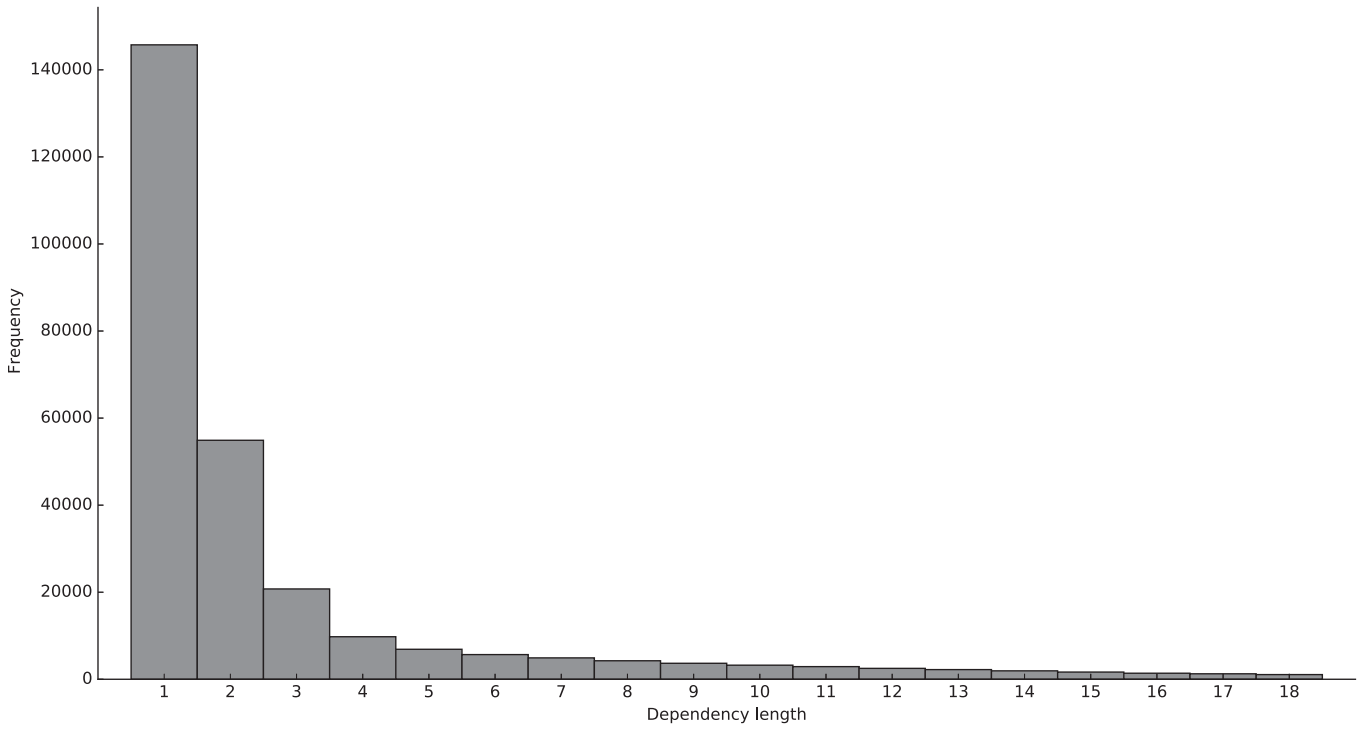


Fig. 1. Dependency length distribution for the Hindi-Urdu Treebank corpus.

2. Background

In this section, we summarize dependency length minimization approaches, which have been very prominent in the literature on both language production and comprehension. Subsequently, we discuss Surprisal Theory. Finally, we provide a background on factors affecting syntactic choice in the Hindi language.

2.1. Dependency length minimization in language processing

Temperley and Gildea (2018) group existing research on dependency length minimization into three classes based on the methodology adopted: (1). Comprehension experiments (2). Corpus studies examining syntactic choice pertaining to one or more *constructions* (3). Computational simulations over *entire grammars*. Our work uses the second methodology mentioned above. The following subsections discuss prior work adhering to each of these methodologies.

2.1.1. Dependency locality theory and comprehension

Dependency Locality Theory (DLT) is a theory of language comprehension (Gibson, 1998, 2000) which seeks to explain the comprehension complexity of syntactic structures. DLT posits two kinds of processing costs associated with language comprehension, *viz.*, Storage cost and Integration cost. Storage cost represents the cost associated with storing syntactic heads required to complete the current input, while integration cost encodes the cost of integrating a word currently being processed with its previously encountered dependents. Notably, DLT predictions have been confirmed using English relative clause structures illustrated below:

(2) (a). The reporter **who attacked** the senator admitted the error (*subject relative clause*)
(b). The reporter **who** the senator **attacked** admitted the error (*object relative clause*)

In the above examples, the verb *attacked* is linked to the relative pronoun *who* and the integration cost is the count of the number of discourse referents (nouns and verbs) occurring between these words. Thus, Example 2a above containing a subject relative clause has an integration cost of 0, while Example 2b having an object relative clause has an integration cost of 1 (on account of the intervening noun *senator*). So DLT predicts greater processing complexity at the verb *attacked*, in the object relative clause in contrast to the subject relative clause.[3] Self-paced reading experiments reveal that this prediction is borne out. In addition to behavioural experiments involving carefully controlled stimuli, researchers have validated DLT predictions (to a certain extent) using more naturalistic stimuli from reading time corpora in English (Demberg & Keller, 2008) and Hindi (Husain et al., 2015). DLT predictions are violated in verb-final languages like German (Konieczny, 2000) and Hindi (Vasishth & Lewis, 2006), leading to *anti-locality* effects (*i.e.*, increase in head-dependent distances causing a speedup at the final verb, contra DLT prediction). Vasishth and Lewis (2006) proffered a unified explanation for both locality and anti-locality effects using the ACT-R framework (Anderson et al., 2004) resorting to the concepts of *decay* and *interference*, themes we take up in Section 4 for detailed discussion.[4]

2.1.2. DLT and syntactic choice

Accessibility-based accounts of language production (Bock & Warren, 1985; Ferreira & Dell, 2000; Prat-Sala & Branigan, 2000) predict that more accessible phrases are realized before relatively less accessible ones. Such accounts assumed short constituents to be more accessible compared to long ones, thus connecting accessibility with length. Accessibility-based accounts thus predict a "short before long" preference across the board *i.e.*, in both preverbal and postverbal domains. Spontaneous production experiments involving English (Arnold et al., 2000; Stallings et al., 1998) validated this prediction, leading to a

[3] The original DLT formulation included empty categories. But Temperley (2007) discuss how these non-lexical items do not affect DLT predictions greatly.

[4] Lewis and Vasishth (2005) is the only computationally implemented model of dependency completion cost to the best of our knowledge.

short-first principle. However, the predominance of preverbal long-short orders in verb-final languages argue against relative length-based definitions of accessibility in language production (Hawkins, 2004; Jaeger & Norcliffe, 2009). In this context, dependency length minimization accounts like DLT and Hawkins' Early Immediate Constituents (Hawkins, 1994; EIC) offer a unified explanation for constituent ordering patterns in both SVO (example English) and SOV (example Japanese) languages. DLT and EIC prefer constituent orders where the overall head-dependent distance is minimized. Preverbally, long constituents followed by short ones result in lower dependency length for the entire sentence compared to short-long orders. The opposite patterns are true in the postverbal domain, *i.e.*, short-long orders contribute to lower overall dependency lengths for sentences than long-short orders. Fig. 2 depicts examples from Hawkins (2004) and illustrates DLT predictions for both English (head-medial language) and Japanese (head-final language). In a nutshell, as illustrated in the figure, head-medial structures display a short-long preference of constituents postverbally (Fig. 2), while head-final structures display the opposite pattern (i.e. long-short order) preverbally (Fig. 2). Subsequent work involving production experiments has built upon Hawkins's EIC principle in order to find psychologically real explanations[5] for constituent ordering patterns in SOV languages like Japanese (Yamashita & Chang, 2001), Korean (Choi, 2007) and Basque (Ros et al., 2015).

2.1.3. Computational simulations

The past decade has seen the emergence of a strand of work investigating whether processing load (quantified using dependency length and surprisal) is a significant factor in the context of the *entire grammar* as opposed to *particular constructions*. Computational simulations over entire grammars (by linearizing dependency trees) have shown for a wide variety of languages that dependency lengths of natural languages are lower than random chance baselines (Futrell et al., 2015; Gildea & Temperley, 2010). Using a similar approach, Gildea and Jaeger (2015) found that both dependency length and *n*-gram surprisal (estimated using word trigram models) of natural language are lower than random for English, German, Arabic, Czech and Mandarin.

2.2. Surprisal theory

The Surprisal Theory of language comprehension was originally proposed by Hale (2001), based on Claude Shannon's (1948) pioneering work on information theory. Levy (2008) elaborated Hale's foundational insight into a theory which posits that probabilistic knowledge (attained from prior linguistic experience) helps comprehenders form expectations about interpretations of previously encountered structure as well as upcoming material. For every word, this theory defines a measure of comprehension difficulty called surprisal, which is correlated with various eye movement measures of reading time (Boston et al., 2008; Demberg & Keller, 2008) as subsequent research attested. More recently, Linzen and Jaeger (2016) provided robust evidence for both surprisal and uncertainty about the full structure (not merely the next prediction step as quantified by uncertainty of the verb complement) in predicting reading times. In this work, we estimated the following per-word surprisal measures defined below:

(1). **n-gram surprisal**: Mathematically, *n*-gram surprisal of the $(i+1)th$ word, w_{i+1}, based on a traditional *n*-gram model is given by $S_{i+1} = -logP(w_{i+1}|w_{i-n+2}, ..., w_{i-1}, w_i)$, as defined by Hale (2001).

(2). **Dependency parser surprisal** was computed using the probabilistic incremental dependency parser developed by Agrawal et al. (2017), based on the parallel-processing variant of the *arc-eager* parsing strategy (Nivre, 2008) proposed by Boston et al. (2011). This parser maintains a set of the *k* most probable parses at each word as it proceeds through the sentence. A maximum-entropy classifier (which encodes an underlying probabilistic grammar *G*) is used to estimate the probability of a transition from one parser state to the next, and the probability of a parser state is taken to be the product of the probabilities of all transitions made to reach that state. This parser can thus be used to define a measure of surprisal: for the i^{th} word in a sentence, we first define the *prefix probability* α_i as the sum of probabilities of the *k* maintained parser states or derivations upto word *i* (denoted by $\{d_{i1}, ..., d_{ik}\}$):

$$\alpha_i = \sum_{j=1}^{k} Prob(d_{ij}) \quad (1)$$

The dependency parser surprisal at word $i+1$ is then computed as:

$$S_{i+1} = -log(\alpha_{i+1}/\alpha_i) \quad (2)$$

The dependency parser surprisal of the $(i+1)^{th}$ word is computed as the negative log-ratio of the sum of probabilities of maintained parser states at word $i+1$ to the same sum at word *i*.

Since our study involved classifying Hindi sentences into reference and variants, we calculated feature values at the sentence level for both reference and variant sentences. Trigram and dependency parser surprisal scores for each sentence were computed by summing the respective surprisal value of each non-punctuation word. For Hindi, Agrawal et al. (2017) showed that lexicalized surprisal derived from their incremental dependency parser predicts first pass reading time in the Potsdam-Allahabad corpus of Hindi text, even after incorporating word-level controls like syllable length, unigram and bigram frequencies in a mixed effects model. However, inspite of such success, surprisal theory has failed several confirmatory tests and the theory requires revision (see Futrell et al., 2020; for a recent work in this direction). In a study of Russian relative clause comprehension, Levy et al. (2013) falsified a key prediction of surprisal. They showed that reading times at relative clause (RC) verbs are highest at non-local conditions where there is an intervening RC NP between relative pronoun and the RC verb and lowest in local conditions where relative pronoun and RC verb are adjacent. Based on a study of German verb-final structures using eye-tracking data, Levy and Keller (2013) argued that in addition to surprisal, the theory must take memory into account. However, their results did not stand the scrutiny of a large scale replication study by Vasishth et al. (2018).

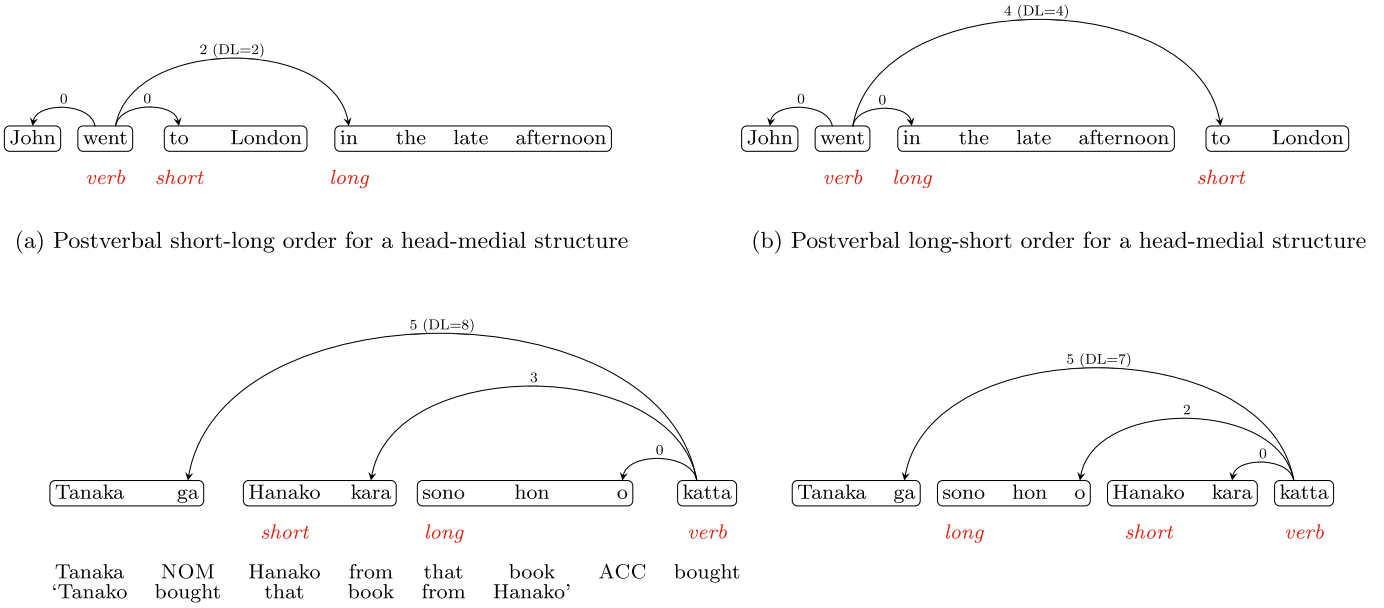
2.3. Factors influencing Hindi constituent ordering

Hindi is a flexible word order language with overt case markers realized as postpositions. Case markers typically encode grammatical functions in language (see Table A.4 in the Appendix for the complete list of Hindi case markers and their grammatical functions). As discussed in the theoretical linguistics literature on Hindi scrambling, Hindi permits movement to both argument (A movement) as well as non-argument (A' movement) positions (Mahajan, 1990). Many authors (Butt & King, 1996; Kidwai, 2000) have also shown that scrambling is influenced by other factors like information status (topic, focus, background and completive information), semantics (definiteness and animacy) as well as prosody (Patil et al., 2008).

3. Methods

In this section we describe an algorithm for creating our dataset (comprising of reference sentences and grammatical variants). Section 3.2 describes a ranking model used to perform the regression and

[5] Liu et al. (2017) provide a comprehensive survey of psychological experiments and corpus studies related to dependency distance.



(a) Postverbal short-long order for a head-medial structure

(b) Postverbal long-short order for a head-medial structure

(c) Preverbal short-long order for a head-final structure

(d) Preverbal long-short order for a head-final structure

Fig. 2. Dependency length and constituent ordering patterns for English head-medial and Japanese head-final structures[6]. (a) Postverbal short-long order for a head-medial structure. (b) Postverbal long-short order for a head-medial structure. (c) Preverbal short-long order for a head-final structure. (d) Preverbal long-short order for a head-final structure.

[6] Overall dependency length (DL) of the structure indicated above each subfigure.

prediction experiments (described in the next section). In that subsection, we also elaborate on the independent and dependent predictors used in our study.

3.1. Variant generation

Our study is based on written text from the Hindi-Urdu Treebank (HUTB) corpus of dependency trees (Bhatt et al., 2009). For each HUTB reference sentence, we created grammatical variants using an algorithm which took as input the dependency tree corresponding to each HUTB reference sentence. Subsequently, the algorithm generated variant sentences by permuting preverbal constituents. Ungrammatical variants were isolated and filtered out automatically using grammar rules extracted from the corpus of HUTB dependency trees. Each HUTB tree consists of head words linked to dependent words via labelled links denoting the grammatical relationship between the words. Table A.3 in the Appendix shows examples of correct and incorrect variants for 3 reference sentences. This setup was inspired from the *over-generate-and-rank* paradigm used in the surface realization component of grammar-based NLG systems (see Rajkumar & White, 2014; and references therein for a comprehensive survey). Fig. 3 depicts an example tree corresponding to Example 3a below, along with a glossary of its grammatical relation labels.

(3) a. [jayalalitha=ne] [is masle par] [pradhaanmantri=ko] [ek patra] likhaa hai (**Reference**)
jayalalitha=ERG this issue PSP prime minister=DAT a letter write.PERF

Jayalalita has written a letter to the prime minister on this issue.

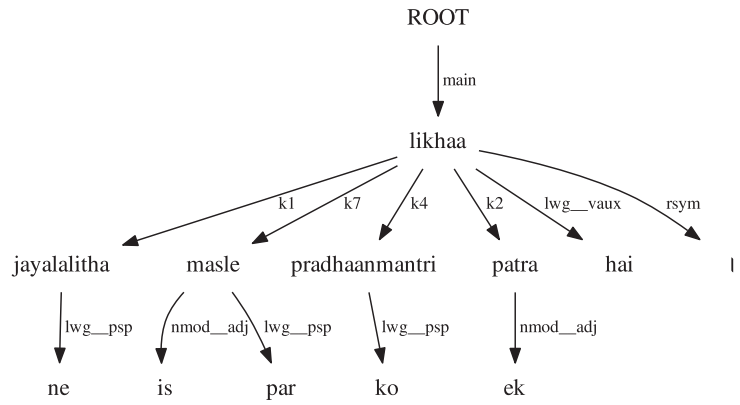
b. jayalalitha ne pradhaanmantri ko ek patra is masle par likhaa hai (*Variant*$_1$)

c. jayalalitha ne ek patra is masle par pradhaanmantri ko likhaa hai (*Variant*$_2$)

From the HUTB corpus, we extracted all the projective dependency trees representing declarative sentences (excluding interrogative and imperative sentence types in the corpus). From this set of trees, we selected all trees whose root node was a finite verb linked to at least two preverbal dependents. Subsequently, we created variants for each reference sentence (as in Example 3a above) using a reordering algorithm which takes as input the dependency tree corresponding to that reference sentence (Fig. 3). The reordering algorithm permuted the preverbal dependents of the root verb (*likhaa* in Fig. 3a) and linearized the resulting tree to obtain variant sentences (Examples 3b and 3c above). In order to automatically ensure that only grammatical variants were chosen, we filtered out variants which did not contain dependency-relation sequences attested in the gold standard corpus of HUTB trees. Such sequences simulate grammar rules used in NLG grammar-based surface realization systems. In the example tree shown in Fig. 3a, *k1-k7*, *k7-k4*, *k4-k2*, *k2-lwg_vaux* and *lwg_vaux rsym* are the root-level dependency-relation pairs (denoting grammar rules). In cases where the number of variants exceeded 100 (a random cutoff),[7] we chose 99 variants randomly. After the aforementioned procedure, we obtained a dataset comprising of 7586 reference sentences and 158891 variants.[8] As shown in Table 1 containing constituent-wise statistics, our dataset consists of sentences with number of preverbal constituents ranging from 2 to 10. As depicted in the table, the majority of data points (46.09%) correspond to sentences of intermediate length containing 5 preverbal constituents.

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

[8] Data can be accessed from the OSF data repository via the link: https://osf.io/zbxau/?view_only=9a67dcc8b66f4df08a65bbf8f8485808.



(a) Dependency tree

Label	Dependency relation
Invariant syntactic relations	
k1	subject/agent
k2	object/patient
k4	recipient
k7	location (elsewhere)
Modifier Relation	
nmod_adj	adjective modifying head noun
Local word group (lwg)	
lwg_psp	postposition
lwg_vaux	auxilliary verb
Symbols	
rsym	symbol relation

(b) Dependency relations

Fig. 3. Example HUTB dependency tree and relation labels. (a) Dependency tree. (b) Dependency relations.

Table 1
Constituent-wise statistics (percentage of constituents in parentheses; ASL is average sentence length in words).

#Constituents	#Data points (%)	ASL
2	1800 (1.13)	19.64
3	11,643 (7.33)	17.49
4	39,352 (24.77)	19.28
5	**73,236 (46.09)**	**21.72**
6	23,976 (15.09)	25.8
7	7184 (4.52)	28.82
8	1403 (0.88)	30.21
9	99 (0.06)	37
10	198 (0.12)	34.5

3.2. Ranking model

In order to test the stated hypothesis, we setup a binary classification task to distinguish the Hindi reference sentence from its artificially generated grammatical variants. The intention was to make a model to choose the corpus sentence from a pair of sentences consisting of the corpus sentence and one grammatical variant of the same (thus a *two-alternative choice* for each reference sentence). Our original dataset described in the previous section contained substantially more variants than reference sentences. Thus to mitigate this imbalance for appropriate classification, we transformed our data set using a technique originally proposed by Joachims (2002) for ranking web pages. For studying syntactic choice in English, Rajkumar et al. (2016) adopted this technique and as they discuss, the transformation converts a binary classification task (labelling a sentence as reference vs variant) into a pairwise ranking task involving the feature vectors of a reference sentence and each of its variants. From the perspective of statistical modelling, the stated transformation facilitates modelling using logistic regression on a balanced dataset as discussed by Joachims (2002).

The overarching cognitive motivation behind this transformation was to model the fact that each reference sentence was generated by the speaker after eliminating a potential grammatical variant. Thus we seek to model the extent to which our independent variables preferred the reference over each variant.[9] We then trained a machine learning model on the difference between the aforementioned feature vectors as per the equations below:

$$\mathrm{w}\cdot\varphi(\mathrm{Reference}) > \mathrm{w}\cdot\varphi(\mathrm{Variant}) \quad (3)$$

$$\mathrm{w}\cdot(\varphi(\mathrm{Reference}) - \varphi(\mathrm{Variant})) > 0 \quad (4)$$

Eq. (3) above shows a data point where the model predicts that the reference sentence outranks one of its variants when the dot product of the feature vector of the reference sentence and **w** (learned feature weights) is greater than the corresponding dot product of the variant sentence. This relationship can also be expressed in the form of Eq. (4), where the feature values of the first member of the pair were subtracted from the corresponding values of the second member. The model determines the choice for a particular referent-variant pair by assessing the sign of the dot product of the learned feature weight with difference of the feature vectors (see Eq. (4)).

We created ordered pairs consisting of the feature vectors of reference-variant sentences. Examples 3a-3b and Examples 3a-3c constitute two such sentence pairs whose feature vectors were paired. Pairs alternate between *reference-variant* (coded as "1") and *variant-reference* (coded as "0"), resulting in a balanced data set that contained either equal number of classification labels of each kind (if the total number of variants is an even number) or a difference of one (if total number of variants is an odd number). Table 2 illustrates the original and transformed feature values for the example sentences referred above. The stated transformation thus enabled us to generalize the effect of the independent variables over a large number of syntactic choice variants associated with a given reference sentence.

We deployed the transformed feature values as predictors in a

Table 2
Joachims' transformation.

(a) Original feature values

Condition	Label	Dependency length	Trigram surprisal	Parser surprisal
Reference	1	16	16.34	0.18
$\mathrm{Variant}_1$	0	18	20.00	0.15
$\mathrm{Variant}_2$	0	17	17.36	0.14

(b) Transformed feature values

Condition	Label	δ dependency length	δ trigram surprisal	δ parser surprisal
$\mathrm{Variant}_1$-Reference	0	2	3.66	-0.03
Reference-$\mathrm{Variant}_2$	1	-1	-1.02	0.04

[9] Another approach would be to compute the percentage of times the reference would be ranked first in comparison to all other variants as in our earlier work (Ranjan et al., 2019). Though our results could easily be upacked in this format and explained, due to space limitations we do not do so.

Generalized Linear Model (GLM) as implemented in the R package. GLMs are standard models used to estimate the probabilities associated with categorical outcomes using logistic regression. We used the following GLM to ascertain whether dependency length is a significant predictor of syntactic choice:[10]

$$\text{choice} \sim \delta\text{ngram surprisal} + \delta\text{parser surprisal} + \delta\text{dependency length} \quad (5)$$

Here *choice* is a binary choice dependent variable (1 stands for the right choice and 0 denotes the wrong choice). The correct choice is where the corpus sentence outranks a variant paired with it and incorrect choice is the opposite situation *i.e.*, the variant outranks the reference sentence. The above model reflects the idea that the reference sentence that did appear in the corpus must have appeared as a consequence of its properties (dependency length and surprisal), and the artificial variants would be less likely to be produced. The difference in predictor values thus models this scenario in the transformed dataset. The cited R package used Maximum likelihood estimation to obtain a GLM using iteratively reweighted least squares for parameter estimation (Baayen, 2008). The values of independent variables were calculated as follows:

1. **Dependency length**: We defined dependency length as the number of intervening words between each head and dependent. The dependency length of a sentence (reference or variant) was calculated by summing each of the head-dependent distances in the corresponding dependency tree, following Gibson's (2000) word-by-word integration cost for relative clauses. In Appendix A, Table A.1 we illustrate dependency calculations for the dependency tree shown in Fig. A.1 (corresponding to the reference sentence shown in Example 4a).
2. ***n*-gram surprisal**: We estimated *n*-gram surprisal using a trigram model (n=3) over words trained on 1 million sentences from the EMILLE corpus[11] (Baker et al., 2002) with Good-Turing discounting using the SRILM toolkit[12] (Stolcke, 2002).
3. **Dependency parser surprisal**: We used the incremental probabilistic dependency parser developed by Agrawal et al. (2017)[13] to estimate dependency parser surprisal (described in Section 2) using a corpus of 12,000 HUTB projective trees. We divided this corpus into 10 sections and models trained on 9 sections were used to get incremental surprisal scores for the remaining section, thus covering the entire corpus. We conjoined case marking postpositions with the preceding head nouns because the treebank annotates nouns with a feature encoding the form of the upcoming case inflection.
4. **Information status (IS) score**: We incorporated an information status (IS) score reflecting *given* vs *new* considerations as a control into our statistical model. We automatically annotated a subset of our sentences containing subject, object (either direct or indirect or both) and verb with an information status score. There is evidence that languages adhere to the *given before new* principle (Clark & Haviland, 1977) by realizing elements which are already salient in the discourse (by previous mention) prior to new content. We estimated givenness using two factors discussed in the extensive literature on this theme, *viz.*, recent mention as well as pronouns. Ferreira and Dell (2000) proposed the *principle of immediate mention* which states that recent mention of given information makes it easily available. Bock and Irwin (1980) provided early evidence that previous mention of a word in the preceding context in the form of lexically identical (or near-identical) content in written text led to such words being mentioned earlier than new elements. Pronouns typically refer to entities already mentioned in the discourse and languages tend to prefer placing pronouns prior to other NPs in a verb-final construction (Kempen & Harbusch, 2008). A subject or an object constituent was assigned *Given* status if it satisfied either of the following 2 conditions: 1. Any content word in the constituent was present in the previous sentence. 2. The head of the constituent was a pronoun. All other constituents were tagged as *New*. For each reference and variant sentence, IS scores were assigned as follows: a) Given-New order = 1 b) New-Given order = -1 c) Other 2 possibilities: Given-Given and New-New = 0. We concede that though such a score is a very rough approximation of IS constraints, to the best of our knowledge, this is the first attempt to quantify such constraints in a statistical modelling framework. The inherent weaknesses of this particular representation of IS constraints should not be held as detracting from theories of information status.

Previous work has quantified dependency length using various measures like number of syntactic nodes (Wasow, 2002), words (Temperley, 2007), intervening nouns and verbs (Gibson, 2000) or prosodic units like number of primary stresses (Anttila et al., 2010). As Szmrecsanyi (2005) notes, all these measures are highly correlated. However, Gibson's work also posits maximal integration cost (*i.e*, maximum value of integration costs at all words in a sentence) to quantify complexity of an entire sentence (Gibson, 1998, 2000). We adopted the summation approach following the reasoning offered in Temperely's (2007) work on DLT and English syntactic choice, whereby complexity can be conceived of as computational effort (reflected in comprehension time). Since each dependency contributes to the integration cost of a word, the total complexity of a sentence is the sum of all head-dependent distances.[14] Rajkumar et al. (2016) showed that for the task of reference sentence prediction (amidst grammatical variants) in English, dependency length defined using number of intervening words, nouns and verbs as well as syllables resulted in almost identical prediction performance. They found that there was no statistically significant difference (as per McNemar's test) between the prediction accuracies of the three dependency length measures mentioned above. So we used the word-based definition of dependency length in this work.

4. Results

This section describes the computational experiments we performed and their results. Section 4.1 summarizes the results of a preliminary corpus study on constituent ordering patterns in the HUTB corpus. Section 4.2 presents the results of our correlation and regression experiments. Section 4.3 provides an overview of the prediction accuracy results of our regression models. Finally, Section 4.4 describes the impact of dependency length and surprisal measures on data points belonging to the conjunct verb construction as well as non-canonical word orders.

4.1. Corpus study

As discussed in Section 2, dependency length minimization accounts predict that a sentence containing postverbal short-long constituent order has a lower dependency length compared to the long-short order. Preverbally, the opposite patterns hold true, *viz.*, long-short orders minimize dependency length compared to short-long orders. In order to test the above predictions on Hindi data, we traversed recursively

[10] We adopted the R GLM format for presenting the model: the dependent variable occurs to the left of '∼' and independent variables occur to the right; δ denotes difference between features.

[11] https://www.lancaster.ac.uk/fass/projects/corpus/emille/

[12] http://www.speech.sri.com/projects/srilm/

[13] https://github.com/samarhusain/IncrementalParser/

[14] As noted by a reviewer, alternate definitions like dissimilarity between sets of individual dependency lengths of reference and variants are also possible for complexity.

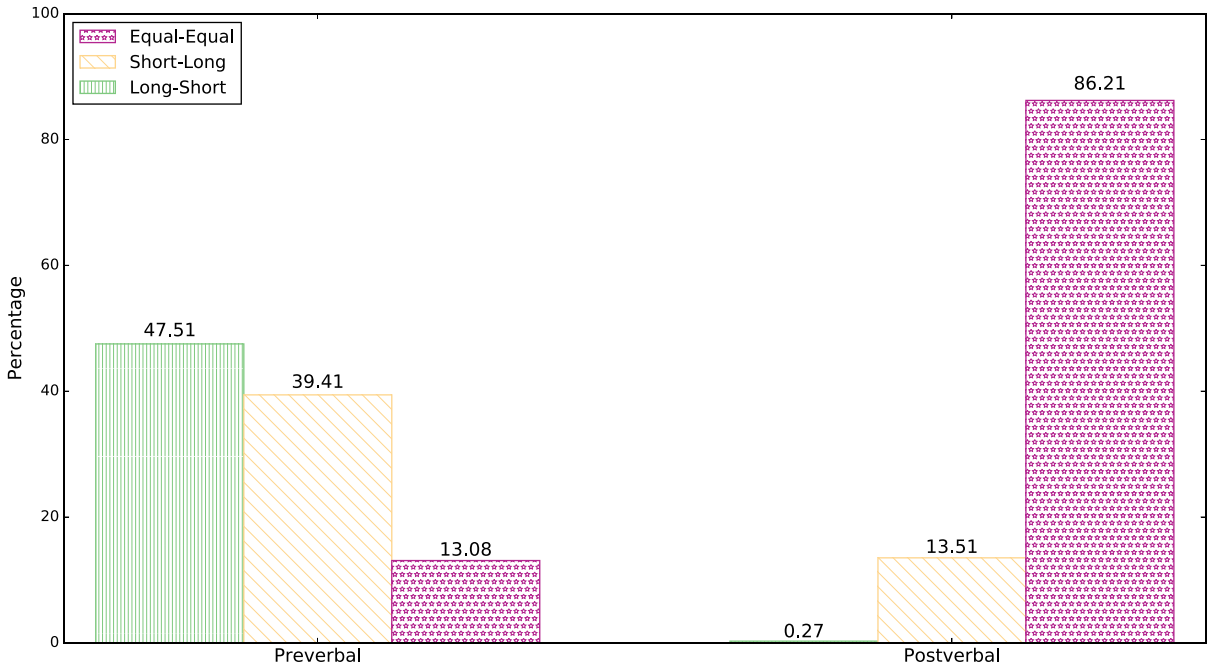


Fig. 4. Hindi preverbal (20,750 pairs) and postverbal (2599 pairs) constituent sequences.

through projective dependency trees corresponding to each reference sentence in our dataset and inferred constituents using heuristics.[15] Subsequently, we computed the number of pairs of adjacent constituent sequences belonging to various types. Fig. 4 depicts our results. Preverbally there is a preponderance of long-short constituent orders over short-long orders, while postverbally, short-long sequences outnumber the long-short sequences (which are few in number). Thus, preverbal Hindi constituent order is influenced by considerations of dependency length minimization. Our results are in line with previous studies discussed earlier for other languages (Hawkins, 2004; Yamashita & Chang, 2001). Our subsequent experiments are solely focussed on word order variation in the preverbal domain as that is the main locus of word order variation in Hindi (see Appendix A for examples). The reversal of the long vs short orders in the postverbal domain is outside the scope of this study as postverbal constituents are very few in number leading to concerns about data sparsity.

4.2. Correlation and regression experiments

In this section, we test our main hypothesis that dependency length predicts preverbal syntactic choice in Hindi in the presence of control variables like trigram and parser surprisal. We tested this hypothesis using the transformed version of our dataset (158891 data points) by introducing trigram surprisal, parser surprisal and dependency length as predictors in a logistic regression model (shown in Eq. (5)) aimed to predict corpus sentences over grammatical variants. The correlation matrix shown in Fig. 5 illustrates that trigram surprisal helps the GLM discriminate between correct and incorrect choices even while occurring in combination with both the other predictors. Moreover, dependency length exhibited low correlation with both trigram and parser surprisal. This result follows the trend reported in previous studies on English sentence comprehension (Demberg & Keller, 2008) as well as syntactic

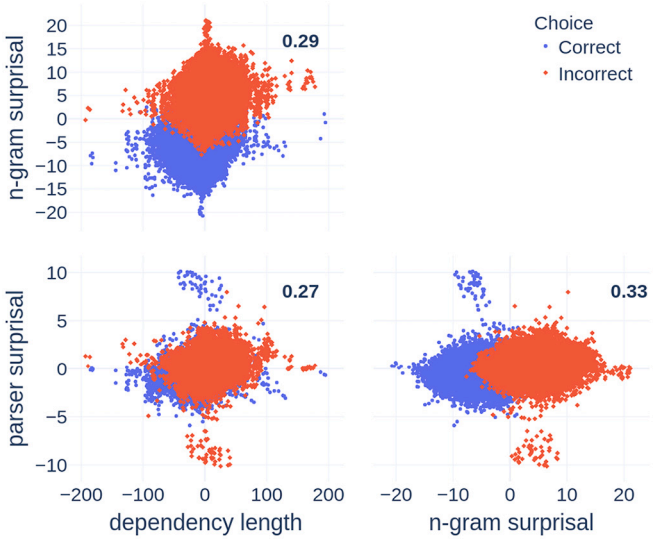


Fig. 5. Scatterplot matrix showing correlations between predictors (Pearson's coefficient of correlation also shown).

choice (Rajkumar et al., 2016). Demberg and Keller (2008) attributed such a low correlation to the complementary nature of integration and surprisal costs.[16] They pointed to the fact that integration cost is a backward-looking cost modelling the integration of a head with a previously encountered dependent, while surprisal is a forward-looking cost reflecting the processing load of an upcoming word given a previous lexical or syntactic context. In our study, parser surprisal and trigram surprisal also demonstrated low correlation with each other. Thus, overall, our correlation results indicate that the three predictors in our study model different parts of the data and a complete account of

[15] Yadav, Vaidya, & Husain, 2017 presents a principled approach for dependency to constituency structure conversion.

[16] While modelling reading times using the Dundee corpus of 10 subjects, Demberg and Keller found that DLT predictions were validated on higher values of dependency length.

Table 3
Regression model containing three predictors (158,891 data points; all predictors except intercept term significant $p < 0.001$).

Predictor/intercept	Estimate	Std. error	z-value
Intercept	-0.003	0.010	-0.247
Trigram surprisal	-1.009	0.006	-164.17
Parser surprisal	-0.799	0.015	-55.05
Dependency length	-0.019	0.0008	-22.23

Table 4a
Prediction performance (158,891 data points; each row refers to a distinct model).

Individual and collective prediction accuracies (*** $p < 0.001$ McNemar's two-tailed significance compared to model on previous row)	
Predictor(s)	Prediction accuracy%
INDIVIDUAL	
Dependency length	60.04
Parser surprisal	69.62***
Trigram surprisal	91.01***
COLLECTIVE	
Trigram + parser surprisal	91.34***
All predictors	91.42***

Table 4b
Relative performance of combinations of predictors

Predictors	%Same (%Correct) prediction
All	38.71 (36.99)
Dependency length + Parser surprisal	55.93 (40.04)
Dependency length + Trigram surprisal	58.11 (54.41)
Trigram surprisal + Parser surprisal	61.00 (57.13)

language processing (both production and comprehension) needs to factor in all three measures, a point emphasized in Rajkumar et al. (2016).

Based on prior work summarized in the introduction and background sections (Sections 1 and 2), our prediction is that the log odds of predicting the reference sentence in a reference-variant pair increases with decreasing values of all the three predictors (reflected by negative sign for their regression coefficients). Essentially, the value of each of the three predictors for a given reference sentence tends to be lower than the corresponding value for any of its variants. For the entire dataset, the boxplots depicted in Fig. A.2 in Appendix A show the mean difference between mean predictor values of variants and reference sentences. As evident from the plots, the variants have higher predictor values as compared to the corpus reference sentence. Now we turn to a discussion of the results of our regression experiments depicted in Table 3. Table A.5 in the Appendix provides the probability of predicting the correct choice (*i.e.* 1) using a model trained using normalized predictor vectors. All three measures are significant for the task of predicting reference sentences and have a negative regression coefficient, validating our original prediction. Thus the log odds of producing the reference sentence increases with the decrease in dependency length as well as suprisal values of the reference sentence with respect to its paired variant. The negative coefficient shows that reference sentences tend to have lower values of dependency length and surprisal compared to grammatical variants. This result confirms our original hypothesis that reference sentences tend to minimize dependency length even in the presence of competing surprisal-based predictors. The negative coefficient of surprisal indicates that reference sentence orders are very predictable.

Moreover, adding dependency length into a model containing trigram and parser surprisal demonstrated the significant effect of dependency length as is evident from a log-likelihood test ($\chi^2 = 505.5$; $p < 0.001$). Vasishth and Lewis (2006) attribute the underlying cognitive processes behind locality effects to time-based *decay* or *interference*. Decay is the loss of activation (exponentially or otherwise) with the distance between words (Anderson & Paulson, 1977). Interference is accrued due to similarity between intervening nouns along the dimensions of meaning, sound or other attributes like concreteness (Van Dyke, 2007; Van Dyke & McElree, 2006). However, decay lacks empirical support as an explanation for forgetting in short-term/working memory tasks (Oberauer & Lewandowsky, 2013), leaving interference as a plausible explanation. Thus the negative sign for the dependency length coefficient suggests proactive interference or low decay, themes which we examine more critically in the next section pertaining to the accuracy of various predictors in predicting the reference sentence (over variants). Overall, our regression results illustrate that Hindi tends to prefer word order choices that minimize both expectation and memory-based comprehension costs.

4.3. Prediction accuracy experiments

In order to quantify the relative contributions of the three predictor variables, we evaluated each model (containing combinations of the three predictors) by calculating its prediction accuracy, which is the percentage of data points where the model made the correct prediction (*i.e.*, chose the reference sentence over the variant paired with it). We divided our entire dataset into 10 distinct sections and models trained on 9 sections were used to generate predictions for the remaining section. Table 4 provides a comprehensive picture of the individual and collective prediction performance of our predictors. The prediction accuracy

of a model containing a single predictor thus represents the degree of preference of that predictor. As illustrated in Table 4a, in terms of individual performance, trigram surprisal accounted for the overwhelming majority of our data (91.01% prediction accuracy), while parser surprisal was the second best predictor (69.62%). Dependency length displayed the lowest individual performance (60.04%). Over a baseline model comprising of trigram and parser surprisal, dependency length induced a small but significant increase of 0.08% in accuracy ($p < 0.001$ using McNemar's two-tailed test).

Table 4b illustrates patterns of relative performance. All three factors predicted the same output in 38.71% of the cases (with 36.99% success). The three predictors had different outputs in around 61.29% of the cases as inferable from the same table (100 minus %Same). Dependency length and each surprisal measure predicted the same output for more than 50% of the cases, while the two surprisal measures had identical predictions for 61.00% of the data (with 57.13% success). Thus our results indicate the overwhelming impact of trigram surprisal on syntactic choice while competing with parser surprisal and dependency length. The efficacy of trigram surprisal stems from the fact that it is able to detect changes at constituent boundaries as illustrated by the following examples (we used dependency structures and bracketing is merely for demonstration purposes):

(4) a. [bhajapa=ne] [*kendr aur keral sarkar-par*] [bharatiy driver em.ar. kutty-ki hatya=ke liye jimmedar taliban=ke sath] [nipatane=mein] [dhilaee baratane=ka aarop] lag-aa-ya hai (68, 64.74, 8.21)
BJP=ERG centre and Kerala government-on Indian driver M.R. Kutty-GEN murder=GEN for responsible Taliban=GEN with deal laxness action=GEN accusation do-CAUS-PFV.M be.PRS.SG

BJP has accused the centre and Kerala governments over laxness in dealing with Taliban, accused in the murder of Indian driver M. R. Kutty.

b. [bhajapa ne] [bharateey draivar em.ar. kutty ki hatya ke liye jimmedar taliban ke sath] [nipatane mein] [*kendr aur keral sarkar par*] dhilaee baratane ka aarop lagaaya hai (63, 67.36, 7.50)

In the reference sentence Example 4a above, when our reordering algorithm moved the second constituent *kendr aur keral sarkar par* (which follows the subject) to a position closer to the final verb, we obtained the variant sentence, Example 4b. The two sentences are identical in all other respects. As evident from the values of the predictors shown above, the reference sentence from the corpus has higher dependency length as well as parser surprisal compared to the variant. However, the trigram surprisal of the reference sentence is lower than that of the variant. In such a situation, the higher magnitude of the trigram predictor's coefficient (refer Table 3) compared to the other two predictors, resulted in the model choosing the reference sentence over the variant.

Since the syntactic surprisal estimates were computed using a dependency parser, they do not directly detect patterns at constituent boundaries and hence may not be as effective in picking up (ir)regularities at these boundaries. Data sparsity might also be a factor impacting the performance of our parser, the number of possible syntactic structures being rather greater than the number of possible trigrams (a point raised by a reviewer). We emphasize the importance of constituent boundaries as variants are created by switching preverbal constituents. Trigrams that lie within constituents will be the same across both reference and variant sentences and hence will not help distinguish between them. Trigrams at constituent boundaries could involve two kinds of cases[17]: 1. Trigrams spread across the boundary between two long constituents with words at the edges acting as cues about the nature of the constituents; 2. Trigrams encompassing one, two, or three short constituents. Chater et al. (2016) proposed an integrated computational model of production and comprehension, where chunk boundaries are successfully detected using changes in the incremental bigram probabilities at each word. This lends credence to the idea that the efficacy of trigram surprisal is at the boundaries of constituents.

Recent works in the comprehension literature have also pointed to effects of trigram surprisal over and above syntax-based language models encoding more context (Demberg & Keller, 2008; Fossum & Levy, 2012). The success of trigram models point to the activity of locally coherent syntactic sequences (with syntactic and pragmatic interpretations) in sentence comprehension (Tabor et al., 2004), which trigram sequences model effectively. Research in English has shown that local parses can induce analyses counter to the global parse and sequence models are effective in modelling such lexical phenomena (Corley & Crocker, 2000; Crocker & Brants, 2000; Tabor et al., 2004). The experiments by Tabor et al. (2004) used stimuli like *The coach smiled at the player tossed the Frisbee*, where the correct global parse attaches *tossed the Frisbee* as a reduced (passive) relative clause modifier of the noun *player*. The fact that the subsequence *player tossed the Frisbee* can be interpreted as an active clause as well is an instance of a diabolical local sequence.[18] In Hindi, there is preliminary evidence that locally coherent sequences can counteract predictions at argument nouns about the final verbal head (Bhatia & Husain, 2018, 2019). We present a more nuanced discussion of this explanation in Section 5.

4.3.1. Impact of dependency parser surprisal

In this section, we seek to evaluate the performance of our dependency parser thoroughly. The motivation for this emerges from the fact that Surprisal Theory explains anti-locality effects in sentence comprehension (discussed in Section 2) using the idea that greater syntactic context confers more information to comprehenders, enabling them to predict the final verb better. More preverbal dependents can result in a speedup at the final verb as more dependents potentially provide sharper surprisal estimates about the identity and location of the final verb to the comprehender (Levy, 2008). To meet the stated objective, we examined the performance of our surprisal estimates derived from an incremental dependency parser on the following subsets of our data:

1. *Locality cases*: Reference sentence has lower dependency length compared to the variant sentence. So dependency length has 100% prediction accuracy for these cases.
2. *Non-locality cases*: Reference sentence has higher dependency length compared to the variant sentence resulting in 0% prediction accuracy for dependency length. Following Rajkumar et al. (2016), in this paper, instead of anti-locality, we use the term *non-locality* as ease of comprehension at the final verb might be accompanied by points of difficulty earlier in the clause.
3. *Zero-locality cases*: Reference sentence has the same dependency length as the variant sentence leading to 0% prediction accuracy for dependency length.

Fig. 6 shows the performance of both trigram and parser surprisal on the locality, non-locality and zero-locality subsets of our data. As the results indicate, trigram surprisal performed much better than parser surprisal in all three splits. The individual performance of parser surprisal is highest for locality cases and lowest for non-locality cases, with zero-locality cases falling in between. Parser surprisal does confer small but significant improvements over trigram surprisal in all the three bins described above (McNemar's test two-tailed significance: $p < 0.001$ for locality and zero-locality bins; $p < 0.01$ for the non-locality bin). In our dataset, we found the following examples involving non-locality

[17] We are indebted to a reviewer for this suggestion about constituent boundaries.

[18] Bicknell et al. (2009) offers evidence that such diabolical sequences exist in naturalistic corpus data as well.

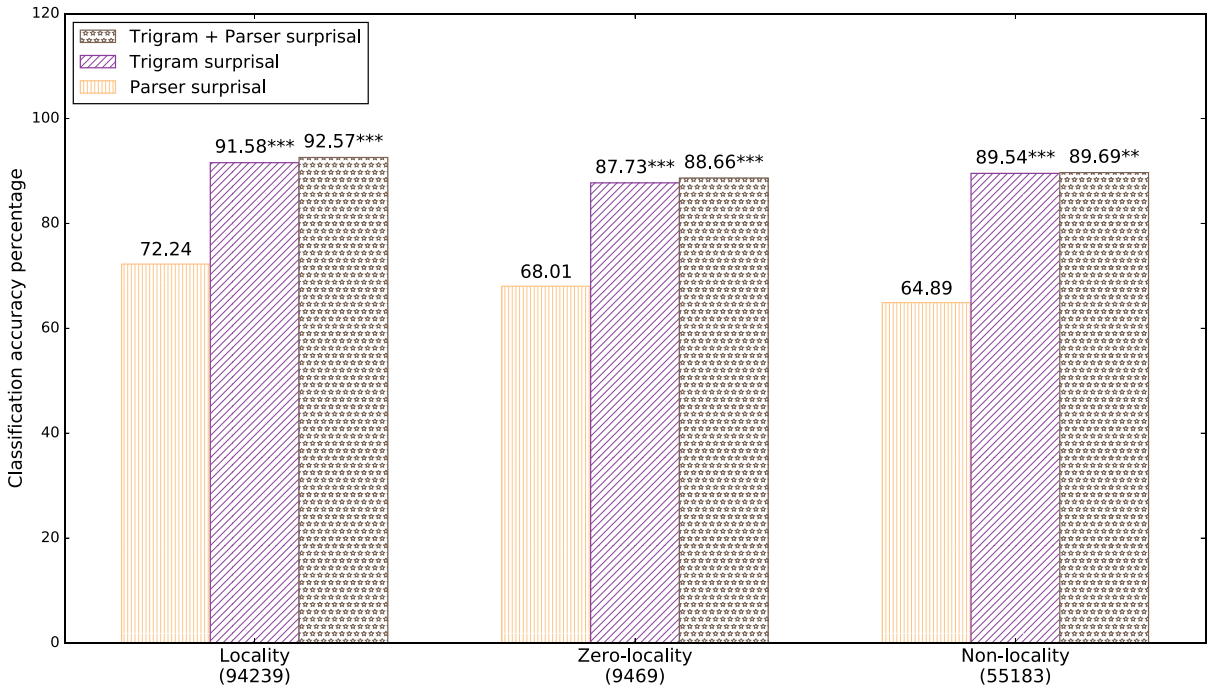


Fig. 6. Prediction accuracy of surprisal (number of cases provided in parentheses; McNemar's two-tailed significance against previous bar indicated using: $***p < 0.001$; $**p < 0.01$).

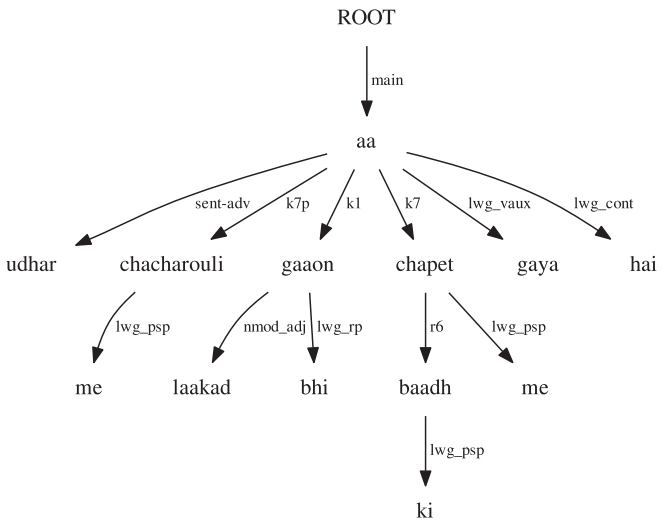


Fig. 7. Example tree depicting identical case inflection *me* marking heads of two preverbal constituents.

illustrating the impact of dependency parser surprisal on preverbal constituent ordering (dependency length, trigram and parser surprisal given alongside):

(5) a. [udhar] [chacharouli=**me**] [laakad gaaon bhi] [baad=ki chapet=**me**] aa ga-ya hai (26, 28.99, 0.66)
there Chacharouli=LOC wood village also flood-GEN hit=LOC come go-PST.M.SG be.PRS.3SG

There, the wooden village in Chacharouli has also been hit by the flood.

b. udhar, laakad gaaon bhi chachrauli **me** baad ki chapet **me** aa gaya hai (25, 27.59, 0.70)

We now offer some hypotheses about why dependency parser surprisal on its own is able to select the reference sentence above, while the other two measures were not effective there. In the reference sentence, Example 5a, the two heads (*chacharouli* and *chapet*) marked by the same locative marker *me* (in bold above) are separated by an intervening constituent (Fig. 7 depicts the tree for this sentence). The variant sentence placed these two constituents together resulting in lower dependency length compared to the reference sentence. The reference sentence above reflects the predominant pattern in the corpus of reference sentences *i.e.* adjacent constituents tend to have non-identical case marking (Ranjan et al., 2019). Mohanan (1994) showed that adjacent nouns with identical case inflections are prohibited, adhering to the Obligatory Contour Principle (OCP) constraint on Hindi word order. Section 5.3 proffers cognitively grounded explanations based on *prediction* and *interference* proposed in the literature to account for the impact of case markers. Both dependency length and trigram surprisal are not able to model this choice correctly. A syntactic language model can effectively model such cases as a trigram model cannot capture relationships beyond its two-word window. The success of parser surprisal arises because it factors in case marker features and also ignores the constituent lengths. Empirical evidence for this claim emerges from Ranjan et al. (2019). They used a similar setup for referent sentence prediction (amongst variants) and showed that an artificial version of Hindi sans case markers resulted in a dip in prediction accuracy of 7% for dependency parser surprisal vis-a-vis 2% for trigram surprisal compared to natural Hindi. A more thorough explanation would need to garner additional evidence pertaining to parser rules and probabilities (a theme for future research). This contrasts with dependency length which prefers to place longer preverbal constituents farther from the verbal head (see placement of 3-word constituent *laakad gaaon bhi* and 2-word constituent *chacharouli me* in above example pair).

4.3.2. Binned prediction accuracy

In this section, we ascertain the performance of each predictor relative to length, motivated by certain distinctive trends related to locality and processing in the literature on English. Rajkumar et al. (2016) point out that for Engish, DLT predictions have been robustly validated for reading experiments containing constructed stimuli (Levy et al.,

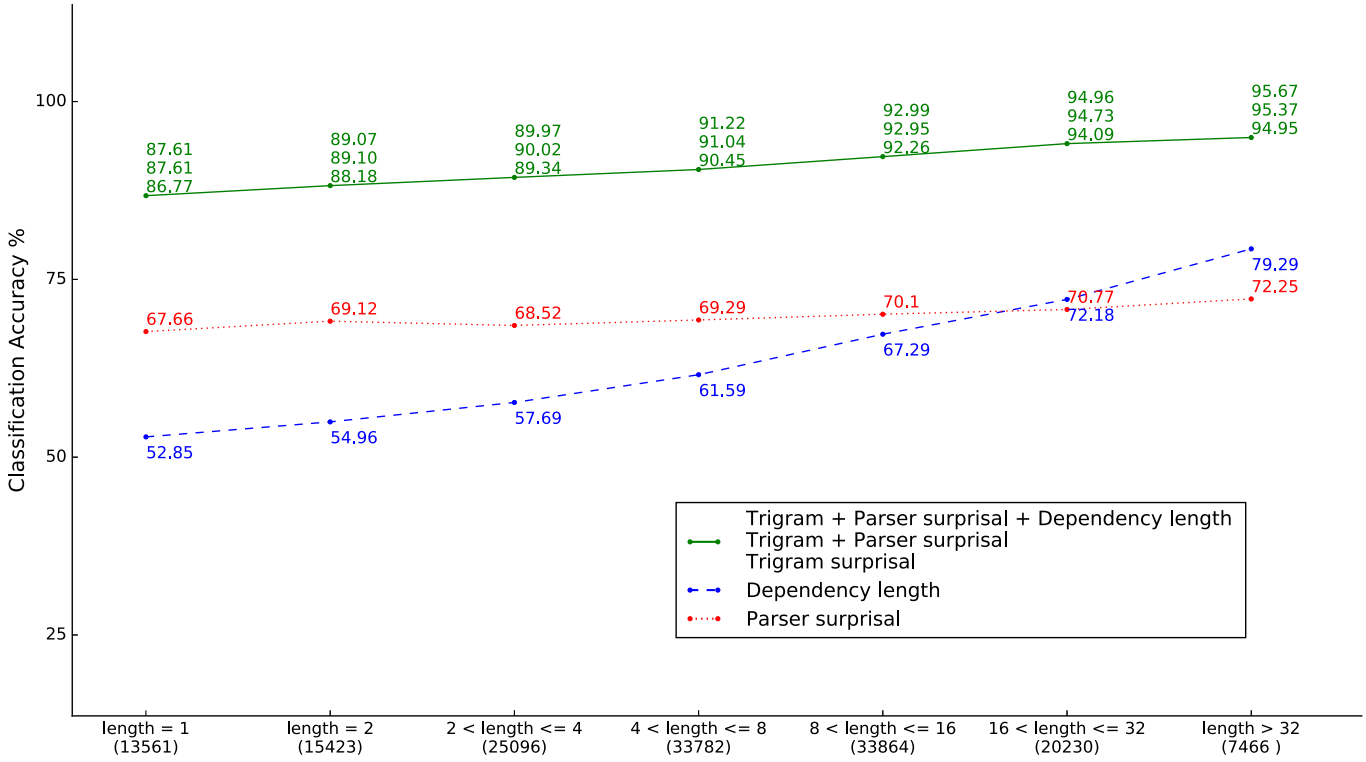


Fig. 8. Classification accuracy for bins of dependency length difference (bin-wise number of data points in parentheses).

2013; Warren & Gibson, 2002). But a weaker effect of locality is seen in broad coverage studies incorporating strong frequency-based controls. Such controls induce reduction or even reversal of the impact of locality (Demberg & Keller, 2008; Shain et al., 2016; van Schijndel et al., 2013; van Schijndel & Schuler, 2013). For example, Demberg and Keller (2008) reported that for English, DLT integration costs resulted in a slowdown in reading times only at higher values. Similarly, in a corpus study investigating English syntactic choice patterns, Rajkumar et al. (2016) showed that the classification accuracies of both PCFG surprisal and dependency length increase steadily with increase in the absolute dependency length difference between reference and variant sentences.

Following the last cited work, we examined the prediction accuracy of all our dependent measures in seven logarithmically sized bins of absolute dependency length difference between reference and variant sentences. Fig. 8 depicts our results. The prediction accuracy of all three measures increased steadily with bin size. Thus, all measures were most effective at the final bin where the dependency length difference between reference and variant sentences was more than 32 words. The relative lack of strength of the dependency length predictor (as is evident from regression coefficients of predictors) is thus offset at high feature values of dependency length. Parser surprisal and dependency length confer small gains over and above trigram surprisal in the last four bins. Our results broadly conform to the trends reported in Rajkumar et al. (2016). However, while in the case of both the English corpora in their study, dependency length prediction accuracy reached the 70% mark in the second bin itself, in Hindi, dependency length attained similar levels only in the penultimate bin (refer Fig. 8). Thus the milder impact of dependency length in Hindi in comparison to English is notable. The figure also shows that trigram surprisal performance ranges from 86 to 94% in all bins. We interpret these patterns theoretically below based on explanations proposed elsewhere in the literature discussed in the previous subsections.

Consistent with prior work discussed in Section 4.2 (Van Dyke, 2007; Van Dyke & McElree, 2006), the efficacy of dependency length at higher bins might be due to proactive interference on account of greater number of dissimilar intervening nouns. In the case of parser surprisal, as discussed before, longer dependencies require more time and context to predict the final verb (Levy, 2008) and hence lead to the greater efficacy of dependency parser surprisal estimates in the case of the latter bins. In the context of HUTB syntactic dependencies, 72.59% of the dependencies shown in Fig. 1 can actually be modelled using either trigrams or bigrams (we connect trigrams to information locality considerations in Section 5.4). These word sequences model locally coherent sequences, which might aid the parsing process (Tabor et al., 2004). They actually recommend larger values of n in word-based language models incorporating more lexical information. Thus higher bins offer more lexical context, resulting in greater prediction accuracy.

Table 5
Construction-wise statistics.

Construction	Reference sentence frequency	Variant frequency
Conjunct verbs	4606	116,020
DO-fronted	133	1663
IO-fronted	101	1353
Overall	7586	158,891

Table 6a
Regression and prediction results for conjunct verb constructions (116,020 data points).

Regression coefficients for model containing all three predictors (all predictors significant, $p < 0.001$)			
Predictor/intercept	Estimate	Std. error	z-value
Intercept	0.001	0.013	0.117
Trigram surprisal	-1.072	0.008	-131.577
Parser surprisal	-0.669	0.017	-39.651
Dependency length	-0.022	0.001	-20.983

Table 6b

Prediction accuracy of distinct models (***McNemar's two-tailed test significance $p < 0.001$ compared to previous row)	
Predictor(s)	Accuracy%
trigram surprisal	92.62
trigram + parser surprisal	93.13***
All predictors	93.23***

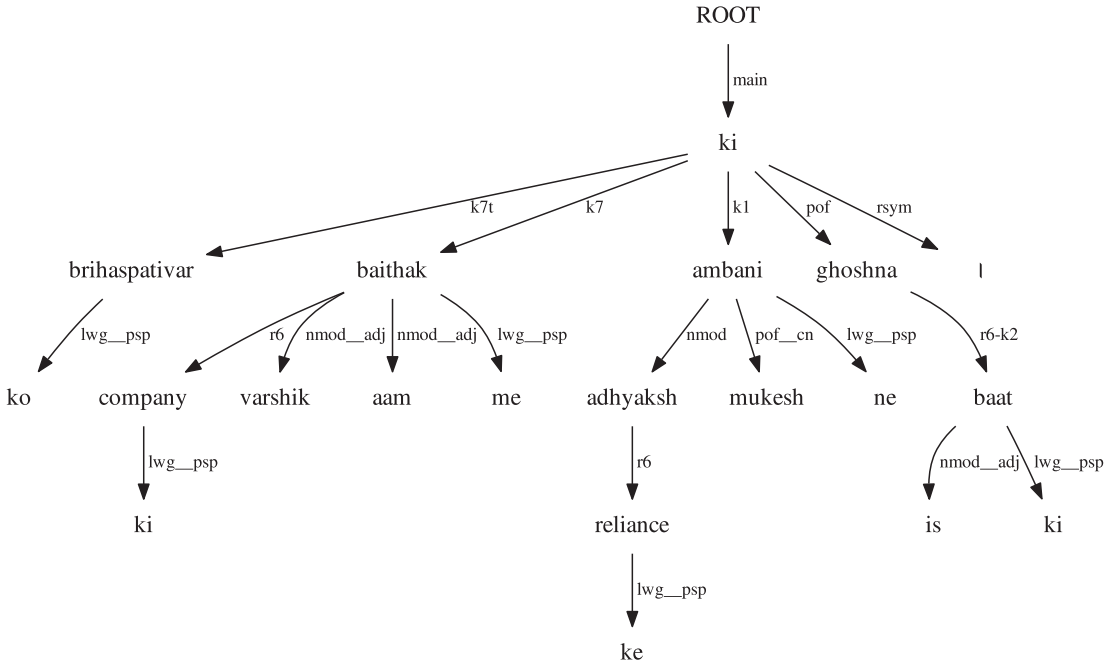


Fig. 9. Example tree depicting a sentence with a conjunct verb.

4.4. *Construction-wise experiments*

Going beyond analyses on our entire dataset, we examined the contribution of each predictor on specific constructions of interest in order to illuminate the predictive power of each predictor. We focus on two constructions, which have been studied in the Hindi sentence comprehension literature, thus facilitating comparison of our results with psychologically real results. First we take up the Hindi conjunct verb construction (Husain et al., 2014) in the next subsection, followed by a discussion of non-canonical word orders (Vasishth, 2004). Table 5 provides a summary of these 2 constructions in our dataset (Table A.2 in the appendix provides a more exhaustive summary of the main constructions in our dataset).

4.4.1. *Conjunct verb construction*

Hindi conjunct verbs are complex predicates comprising of a noun followed by a verb (Kachru, 1982, 2006). We examined the impact of dependency length and surprisal measures in a subset of the data consisting of conjunct verbs (small set marked in the HUTB using the *pof* dependency relation label). In prior work on Hindi sentence comprehension, Husain et al. (2014) examined both non-compositional predicates like *khayaal rakhna* ('care keep/put'; 'to take care of') and compositional predicates like *guitar rakhna* ('guitar keep/put'; 'to put down or keep a guitar'). In compositional predicates, the meaning of the predicate is derived by combining the meanings of its parts. The final verb in non-compositional predicates tends to be more predictable given the noun compared to compositional predicates (ascertained from probabilities resulting from a separate sentence completion task). In the case of compositional predicates, increased distance due to an intervening adverbial phrase between the noun and the verb of the predicate led to a slowdown at the final verb compared to the condition where both parts of the predicate were adjacent. Using a self-paced reading experiment, Husain and colleagues showed that in non-compositional sequences, increasing the noun-verb distance by introducing intervening adverbials did not result in significance differences in reading times at the final verb compared to the condition where both parts of the

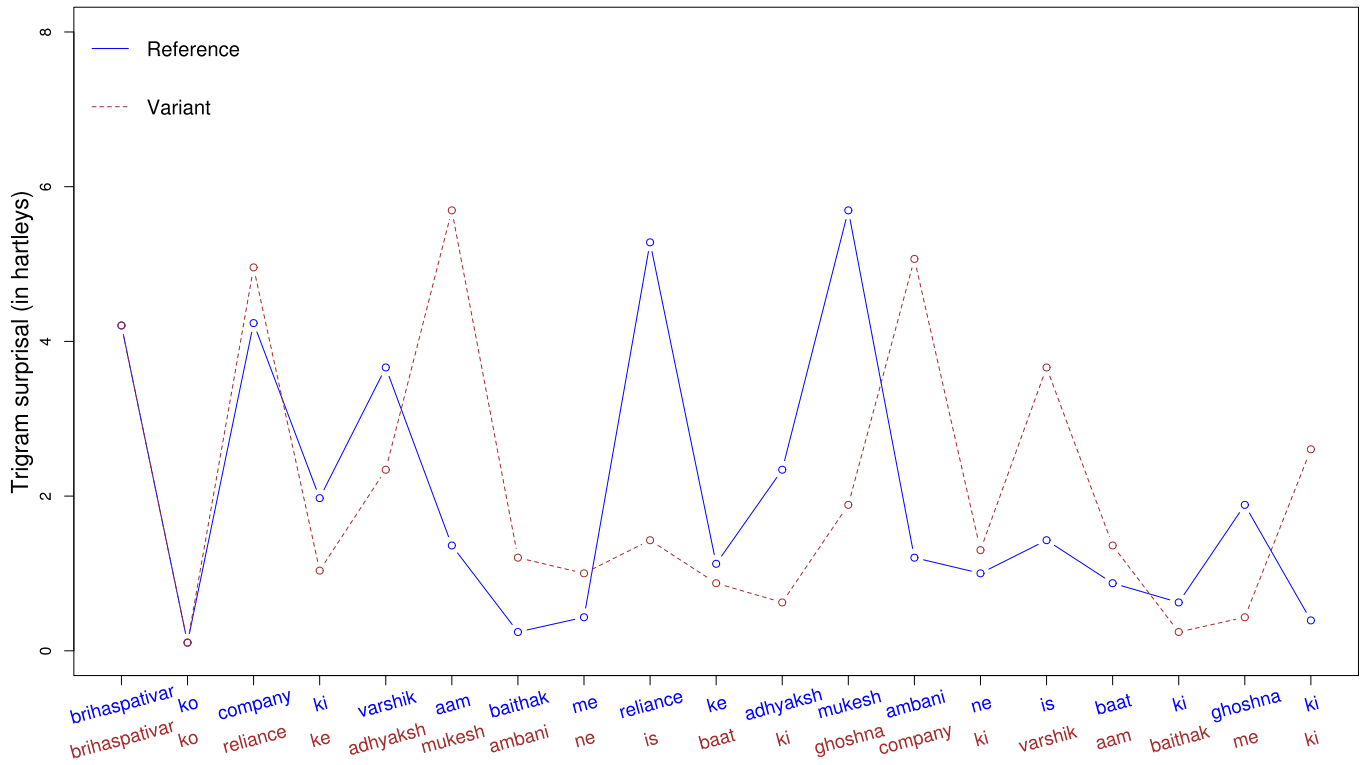


Fig. 10. Trigram surprisal profile.

predicate were together. Thus they showed that locality effects emerged in predicates where the final verb was not predictable.

Our regression results documented in Table 6a show that the log odds of predicting the reference sentence increases with decreasing values of all the three predictors (*viz*, dependency length and the two surprisal measures). Both parser surprisal and dependency length contribute to a significant increase in prediction accuracy over the trigram surprisal baseline (Table 6b). Inspired by the work discussed above, we isolated a set of sentences (one reference sentence and all its 23 variants) containing a conjunct verb *ghoshna ki* ('announced' or 'made an announcement') consisting of a predicating noun *ghoshna* and a light verb *ki*. Fig. 9 depicts the dependency tree of the reference sentence and both parts of the conjunct verb are shown in bold in the example sentences below. For data points corresponding to these examples, our final model (containing trigram surprisal, parser surprisal and dependency length predictors) had a prediction accuracy of 78.26% (correct prediction in 18 out of 23 cases containing intervening material between parts of the conjunct verb). Example 10 in the Appendix illustrates such variants of the reference sentence used in this example. Examples 6a and 6b below illustrate an instance where all the three predictors correctly predicted the reference sentence (dependency length, trigram surprisal and parser surprisal indicated alongside each sentence):

(6) a. [brihaspativar=ko] [company=ki vaarshik aam baithak=me] [reliance=ke adhyaksh mukesh ambani=ne] is baat=ki **ghoshana ki** (40, 38.65, 5.19)
Thursday=at company-GEN annual general meeting=LOC reliance-GEN chairman Mukesh Ambani=ERG this matter=GEN announcement do-PFV.F

On Thursday, the chairman of Reliance, Mukesh Ambani, announced this matter at the company's annual general meeting.

b. brihaspativar ko reliance ke adhyaksh mukesh ambani ne is baat ki **ghoshna** company **ki** vaarshik aam baithak me **ki** (42, 41.11, 5.35)

c. brihaspativar ko company ki vaarshik aam baithak me is baat ki **ghoshna** reliance ke adhyaksh mukesh ambani ne **ki** (42, 41.50, 4.75)

The reference sentence, Example 6a, had lower surprisal (both trigram and parser) and dependency length compared to the variant in Example 6b. In the variant, a long intervening adjunct, *company ki vaarshik aam baithak me* ("in the annual general body meeting of the company"), between *ghoshna* and *ki*, led to greater dependency length compared to the reference sentence. Figs. 10 and 11 depict per-word trigram and parser surprisal profiles respectively of the sentences above. The plots indicate that both trigram and parser surprisal values at the final verb *ki* were lower for the reference sentence in comparison to the variant sentence (high surprisal reflects low predictability). Though *ghoshna* predicts *ki* strongly, in the variant, the intervening element between *ghoshna* and the verb *ki* has a genitive feature *ki* (homophonous with the verb) and hence the parsing model confuses the genitive and the verb. So the verb which comes after the genitive feature has higher surprisal. Thus in addition to increasing the overall dependency length of the variant, the intervening adjunct affected the prediction strength of the predicating noun *ghoshna*, leading to processing difficulty on account of both expectation and locality-based factors at the final verb *ki* as discussed in Husain et al. (2014). However, Example 6c illustrates a variant where the constituent *reliance ke adhyaksh mukesh ambani ne* (without any homophonous *ki*) intervenes between *ghoshna* and *ki*. Here the dependency parser surprisal measure prefers the variant over the reference sentence, pointing to detailed investigations on the nature of intervening elements. Overall, our surprisal measures successfully modelled the fact that nominal and verbal elements of conjunct verbs were adjacent in 92.89% of the cases. We also provide a contrast set of low-expectation counterparts below[19], where in both reference and variant sentences, the final verb *ki* has a higher surprisal (same in both the sentences below) compared to the earlier reference-variant pairs in

[19] We are indebted to Prof. Shravan Vasishth for this suggestion which facilitates a comparison with Husain et al. (2014) and Safavi et al. (2016).

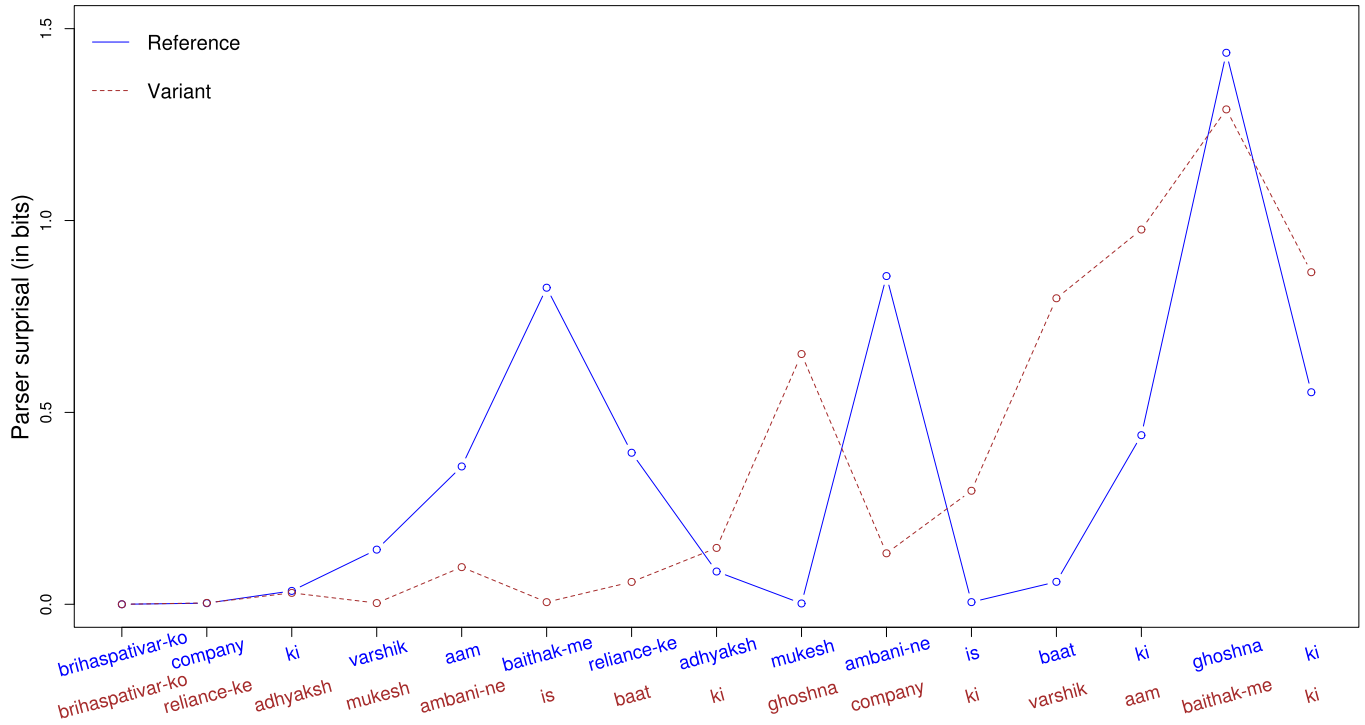


Fig. 11. Dependency parser surprisal profile (case markers are excluded as they are represented as a feature on the preceding head noun during parsing).

Table 7
Regression and prediction results for overall (72,833 points), direct object (DO; 1663 points) and indirect object (IO; 1353 points) fronted cases.

(a) Regression coefficients for overall, DO, and IO models containing all predictors (*** $p < 0.001$)

Predictor/intercept	Estimate	Std. error	z-value
Overall			
Intercept	-0.008	0.015	-0.543
Trigram surprisal	-0.997	0.009	-108.582***
Parser surprisal	-0.707	0.019	-36.532***
IS score	0.284	0.020	13.862***
Deplen	-0.019	0.001	-15.297***
DO			
Intercept	-0.005	0.073	-0.068
Trigram surprisal	-0.529	0.031	-17.324***
Parser surprisal	-0.499	0.124	-4.018***
IS score	0.354	0.066	5.337***
Deplen	-0.037	0.006	-6.127***
IO			
Intercept	0.1180	0.097	1.220
Trigram surprisal	-0.993	0.059	-16.928***
Parser surprisal	-0.792	0.180	-4.394***
IS score	0.148	0.084	1.763.
Deplen	-0.004	0.008	-0.453

(b) Prediction accuracy of distinct models (McNemar's two-tailed test significance compared to previous row: *** $p < 0.001$; ** $p < 0.05$)

Predictor(s)	Overall	DO-fronted cases	IO-fronted cases
Trigram surprisal	91.14	78.77	87.29
Trigram + parser surprisal	91.66***	80.22**	87.66
Trigram + parser surprisal + IS score	91.72**	80.64	87.95
All predictors	91.81**	82.08***	87.66

Table 8
Constituent length statistics in HUTB reference sentences.

Order	Mean length of object	Mean length of subject	# cases
S &DO overall	4.43	3.63	1617
S DO	4.28	3.64	1484
DO S	6.13	3.5	133
S &IO overall	3.92	3.11	769
S IO	3.9	2.95	668
IO S	4.08	4.17	101

Example 6.

(7) a. Taro Aso=ne budhavaar=ko raashtrapati APJ Abdul Kalaam vaanijya mantri Kamalnath aur raashtriya suraksha salaahakaar M K Narayanan=se bhi **mulaakaat ki** (66, 52.02, 5.28)
Taro Aso=ERG Wednesday=ACC President APJ Abdul Kalam commerce minister Kamalnath and national security advisor M K Narayan=ACC also meet do-PFV.F

Taro Aso also met President APJ Abdul Kalam, commerce minister Kamalnath and national security advisor MK Narayanan on Wednesday.

b. budhavaar ko raashtrapati APJ Abdul Kalaam vaanijya mantri Kamalnath aur raashtriya suraksha salaahakaar M K Narayanan se bhi **mulaakaat** Taro Aso ne **ki** (55, 56.35, 6.08)

Here the variant obtained by inserting the subject constituent between the two parts of the conjunct verb *mulaakaat ki* has a lower dependency length compared to the reference sentence, but higher surprisal values. So the model chooses the reference sentence as both surprisal measures dominate the decision. Further experiments are required to ascertain whether these results validate the findings of Husain et al. (2014), especially as lexical surprisal measures are not factored in that work. In contrast to the above work, in a study of Persian noun-verb predicates, Safavi et al. (2016) reported that dependency locality was a robust predictor in both high expectation as well low expectation conditions described above. Both these studies do not factor in the nature of intervening elements (Futrell, 2019) and so further investigation is required.

4.4.2. Canonical and non-canonical word orders

Hindi has a canonical word order of Subject-Indirect Object-Direct Object-Verb (Mohanan & Mohanan, 1994). Comrie (1981) defined canonical word order as the statistically most frequent order, which is also the neutral order in a given discourse context. The comprehension of non-canonical word orders has been investigated in the Hindi sentence comprehension literature (Choudhary, 2011; Mishra et al., 2011; Vasishth, 2004). In addition, works on Hindi and other verb-final languages have revealed role of information status considerations in preverbal constituent ordering (Butt & King, 1996; Ferreira & Yoshita, 2003). Extensive work in English has shown that newness in the discourse is a factor which is independent of constituent length in ordering choices (Bresnan et al., 2007; Gallo et al., 2008; Snider, 2009). Based on Butt and King (1996), who delineated subjects and object constituents as being crucial to information structure, we calculated the Information Status (IS) score in a subset of our data (72,833 points) where both subject and object (either direct or indirect) constituents are present (see earlier Section 3.2 for exact details). This setup thus allowed a controlled study of subject and object alternations in our Hindi data. Our overall results in this subset of the data and performance on direct and indirect object fronting cases therein are shown in Table 7. The positive coefficient of the IS score also reveals that the odds of the reference having *Given-New* order of subject/object constituents is higher than a variant sentence (Table 7a first block). So the IS score contributes towards a significant increase in predicting the reference sentence over and above trigram and parser surprisal (Table 7b first block). The prediction results also show that dependency length induces a significant increase over all the other 3 predictors.

Vasishth (2004) investigated the impact of locality on non-canonical word orders (direct and indirect object fronting) in single center-embeddings in salient as well as non-salient contexts. Vasishth demonstrated that increased distance due to direct object fronting resulted in increased self-paced reading times at the innermost verb compared to the canonical order in both salient and non-salient contexts. In contrast, salient context mitigated the processing difficulty due to distance in indirect object fronted sentences. The subsequent Event Related Potential (ERP) study conducted by Choudhary (2011) also revealed the processing difficulties associated with OSV order. Inspired by Vasishth's work described above, we examined the performance of dependency length and surprisal in predicting reference sentences with non-canonical object-fronted structures. For this purpose, we isolated the following reference-variant pairs of sentences in our dataset: (1). Direct object (DO) is fronted in the reference sentence, while the variant has the canonical order of subject preceding DO. (2). Indirect object (IO) is fronted in the reference sentence, while the variant has the canonical order of subject preceding IO. We used the following HUTB dependency relations connecting the main verb to its dependents to isolate the above subsets: *k2*-DO, *k4*-IO and *k1*-subject. Our regression and prediction results are documented in Table 7. These subsets constitute a very small fraction of our dataset, revealing how infrequent such inversions are.

We deployed construction-specific regression models with trigram surprisal, parser surprisal, IS Score and dependency length as independent variables (introduced in that order) to predict the reference sentence in each of the above subsets of our data. Blocks two and three of Table 7a depicts our regression results. In a model trained on the DO-fronted subset, all the predictors except IS Score showed negative regression coefficients. In the IO-fronted subset, surprisal and dependency predictors showed a negative regression coefficient, but only the two surprisal measures were significant, while dependency length and IS score were not significant predictors of syntactic choice. For studying prediction accuracy, we examined the output of our jack-knifed regression models trained on the subject-object dataset *i.e.*, models trained on 9 folds of this dataset were used for prediction in the remaining fold. Table 7b shows the prediction performance of models containing various features on the IO and DO-fronted subsets of our data. In the case of the IO-construction, none of the other three predictors contributed to significant increases in the prediction accuracy over and above trigram surprisal (significance estimated using McNemar's test). The lack of a locality effect mirrors findings in Hindi sentence comprehension, where Vasishth (2004) showed how discourse context can compensate the processing difficulty induced by indirect object fronting. In the case of direct object fronting, parser surprisal induced a significant improvement in prediction accuracy over trigram surprisal (McNemar's two-tailed significance $p < 0.001$). However, the IS score contributed to a small (non-significant) increase in prediction accuracy over a baseline model comprising of both trigram and parser surprisal. Dependency length further induced a significant increase over and above the other three predictors (McNemar's two-tailed significance $p < 0.001$). The following pair of examples from our dataset illustrate the success of dependency length in predicting DO-fronted non-canonical orders (dependency length, trigram surprisal, parser surprisal and IS score given alongside):

(8) a. [raashtradhvaj tiranga chhape paridhaanon=ke istemaal=par rok hatane sambandhi prastaav=**ko**] [cabinet=**ne**] [aaj] [manzooree] de d-i (14, 53.42, 2.85, 0)
national flag tricolour printed garments=GEN usage=on stop remove related proposal=ACC cabinet=ERG today approval give give-PVF.F.SG

The cabinet approved the proposal to remove the ban on the use of garments having the tricolor (national flag) printed on them.

b. cabinet **ne** raashtradhvaj tiranga chhape paridhaanon ke istemaal par rok hatane sambandhi prastaav **ko** aaj manzooree de di (26, 52.37, 2.79, 0)

The reference sentence above exhibits fronting of the long object *raashtradhvaj tiranga chhape paridhaanon ke istemaal par rok hatane sambandhi prastaav ko* ('the proposal to ban the use of national flag tricolor printed garments') such that it precedes the subject of the sentence *cabinet*. Thus, object fronting led to an overall dependency length of 14 words for the reference sentence, while the variant sentence having the canonical order had a dependency length of 26. The individual strength of dependency length was responsible for ensuring that the model chose the reference sentence over the variant. Thus we show that direct object fronting is preferred when there is an advantage in terms of dependency length minimization over the canonical order. Dependency length is not a significant predictor of indirect object fronting, however. The differential impact of dependency length in direct and indirect object fronting can be accounted by the fact that the majority of direct object fronted cases occur in intermediate and high bins of absolute dependency length difference between reference and variants (average difference is 13.92 words for DO-subset). In contrast, indirect object fronted sentences occur in low and intermediate bin sizes (7.77 words for IO-subset). As discussed before, all measures are effective at higher bin ranges. The lack of any dependency length minimization advantage in indirect object fronting cases can be accounted by the fact that the difference between the length of subject and object constituents is lower for fronted indirect objects in comparison to fronted direct objects (see Table 8). So the long-short preverbal ordering of constituents shown in Fig. 4 is seen only for DO-fronted orders.

Both trigram and parser surprisal were not able to correctly predict the HUTB reference sentence having the non-canonical order in Example 8a. We attribute the failure of surprisal measures to the fact that the reference sentence has the non-canonical accusative-ergative order of case markers associated with the heads of the first two constituents (*ko-ne* markers shown in bold above). The ergative-accusative order (*ne-ko*) is the dominant pattern in the HUTB corpus and the opposite non-canonical order of case markers is seen only in 6% of the total number of 175 cases (Agrawal et al., 2017; Husain et al., 2013). In the reference sentence, *ko* and *ne* markers fall within a trigram window leading to a higher trigram surprisal score reflecting a dispreferred sequence of words compared to the variant sentences where these markers are far apart. Such locally incoherent syntactic sequences can potentially disrupt the parsing process. Parser surprisal estimates are also sensitive to such biases as the parser uses a morphologically rich feature set consisting of word form, lemma, POS tag, tense-aspect-modality annotations and case marker features (Agrawal et al., 2017). Thus the reference sentence gets a higher parser surprisal score. Parser surprisal is ineffective for fronted indirect objects compared to direct objects as 89.2% of indirect objects are case marked in contrast to 29.3% of direct objects (as inferred from a corpus of study of HUTB reference sentences). This suggests that IO-fronted structures have greater tendency to result in dispreferred non-canonical case sequences as compared to DO-fronted structures. Thus, in combination with the case marker of the subject, non-canonical IO-fronted orders (with case marked IO) in reference sentences are dispreferred (over their canonical variants) more often than non-canonical reference sentence orders involving non-case marked fronted DOs. In Section 5 we discuss a cognitively motivated explanation for case marker effects.

5. Discussion

Overall, our results indicate the dominance of lexical expectation effects over locality effects in Hindi syntactic choice. We show that locality considerations are respected in rare, non-canonical structures like fronting direct objects prior to the subject. Prior work in Hindi sentence comprehension attests to the impact of dependency locality on increased reading times at the final verb for the object-fronting construction (Vasishth, 2004). More generally, experimental findings from verb-final languages have revealed a complex pattern of interaction between locality and expectation effects in sentence comprehension.[20] Vasishth et al. (2010) put forth the idea that locality constraints are germane for verb-medial languages like English and Russian, while expectation effects dominate in verb-final languages like German, Japanese, and Hindi. However, subsequent work has shown the existence of both expectation and locality effects in German (Levy & Keller, 2013) and Hindi (Husain et al., 2014). Our findings for syntactic choice lend further credence to the above conjecture articulated in the literature. In the following subsections, we discuss the implications of our findings in light of the low impact of dependency length, case markers, the information locality hypothesis, and language production.

5.1. Low impact of dependency locality

In this section, we explore possible reasons for the relatively weak impact of locality in Hindi constituent ordering by discussing some claims in the literature. In simulations involving entire grammars of English and German (see Section 2), Gildea and Temperley (2010) reported that German does not exhibit a strong tendency to minimize dependency length compared to English. They attributed the weak effect of dependency length in German to the following factors: (1). Crowding and maximal dependency length situation; (2). Case markers; (3). Pronoun placement; (4). Information status considerations. In this section, we discuss cognitively motivated reasons for dependency length not being effective in our study. In the next section, we discuss the cognitive impact of case markers. In this study we have made an initial attempt to incorporate pronouns and *Given-New* considerations as factors in our statistical model. However, we take the view that discourse factors as proposed in the literature constitute a separate level of analysis vis-à-vis cognitive factors, and that they can be translated into cognitive factors as suggested by Arnold et al. (2013) in a comprehensive survey of information structure in language.

Our most significant finding is that dependency length has a weak impact in predicting reference sentences over grammatical variants, as evinced from prediction accuracy over and above a baseline model containing trigram and dependency parser surprisal predictors. Our subsequent analyses revealed that the prediction accuracy of dependency length increases with increase in the absolute dependency length difference between reference and variant sentences. This result follows the trends reported for English in the case of both syntactic choice (Rajkumar et al., 2016) as well as comprehension (Demberg & Keller, 2008). In both situations, only high values of dependency length difference had an impact over and above surprisal-based controls. Thus it is plausible that only very long dependencies induce memory loads strong enough to create an impact on syntactic choice cross-linguistically. Revealingly, exponential (or similar) decay of activation has also long been known to underlie ease of retrieval from memory (Anderson & Paulson, 1977). However, recent work in cognitive psychology suggests that decay has no robust evidence supporting it (Oberauer & Lewandowsky, 2013), and the simulations in Engelmann et al. (2019) corroborate that conclusion for reading time data in sentence comprehension.[21] Future work also needs to investigate whether proactive interference due to intervening nouns is a plausible explanation for this effect (Engelmann et al., 2019; Jager et al., 2017; Vasishth et al., 2019). Thus the overall low impact of dependency length for the entire data might be on account of the underlying cognitive construct behind it, *decay*, not having much relevance for sentence processing. Moreover, the underlying mechanisms probably involve other drivers of comprehension difficulty, like interference and cue-based retrieval cost

[20] Refer to Husain et al. (2014) for a thorough summary of locality and expectation effects in various languages.

[21] For more information on memory retrieval during sentence comprehension, see Lewis et al. (2006), Nicenboim and Vasishth (2018), Vasishth et al. (2019) and Qian and Jaeger (2012) for production.

as evinced in the large sample study pertaining to agreement attraction and reflexive data by Jager et al. (2020). DLT would not be able to explain the dependency completion effects reported in that paper as well as other related work on dependency completion (Engelmann et al., 2019; Jager et al., 2017; Vasishth et al., 2019).[22]

5.2. Success of lexical surprisal

We attribute our trigram model results to increased accessibility of multi-word sequences (Reali & Christiansen, 2007) and local coherence effects (Tabor et al., 2004). Hence, based on evidence from the literature, we hypothesize that reference sentences are more *accessible* compared to variants. The accessibility of words and concepts depend on their frequency, recency of mention, and predictability in the given context (Arnold, 2011). High frequency words are more accessible compared to low frequency words. Highly predictable entities can feature in many conceptual relations resulting in more retrieval pathways and hence are retrieved faster in comparison to entities with low predictability (Arnold, 2011; Bock & Warren, 1985). Jaeger (2011) summarizes evidence to the effect that more accessible elements are available in memory for processing and hence easily retrievable compared to their less accessible counterparts. Further, Reali and Christiansen (2007) found that after controlling for the pronoun-verb frequency, the frequency of the verb alone did not significantly predict reading times. This result suggests that rather than access to individual words, access to word chunk representations plays a role in comprehension. Since surprisal is actually an index of predictability, the effects of trigram and parser surprisal need to be teased apart in terms of their contribution to accessibility in future work.

The local coherence explanation merits a more detailed discussion. The sense of local coherence used by Tabor et al. (2004) pertains to artificially introduced local lexical information driving the parsing process, even if it conflicts with global syntactic analyses. Another sense in which local sequences go counter to the main parse is by modelling naturally occurring dependencies not present in the grammar.[23] Essentially, trigram sequences facilitate processing in instances where the grammar doesn't model local semantic linkages between words (while these may strongly influence processing by native speakers). We present preliminary evidence towards the second sense of local coherence discussed above. Overall our test set contains 248391 trigrams, and 85.2% of these contain at least one syntactic dependency (*i.e.*, syntactic relationship between head and dependent words as defined by a dependency grammar of the language). Essentially, in these cases, both the head and the dependent words are contained within the trigram itself. Notably, the remaining 14.8% of trigrams involve words with no syntactic dependency amongst them, and future studies need to ascertain whether these effectively model cases where the grammar is inadequate. In the reference sentence in Example 4a, the trigram *nipatane mein dhilaee* does not involve a dependency between the words *nipatane* and *dhilaee*. Thus syntactic surprisal estimates from a dependency parser do not model this linkage. In order to test the performance of a constituency parser in modelling such sequences, we also trained an incremental latent variable PCFG parsing model using the Berkeley parser (Petrov et al., 2006), based on a richer split-merge grammar. Syntactic surprisal estimates of that parser predict the reference sentence above over the variant (*nipatane mein dhilaee* falls under a single constituent in the tree corresponding to the reference sentence). Overall, the PCFG parser surprisal is close to (but still inferior to) trigram surprisal in terms of reference sentence prediction accuracy, an observation which we hope to investigate in detail in future work. Further exploration is required to validate the conjecture that trigrams constitute instances of local coherence which facilitate the processor where the grammar fails.

Further inquiries using sentence comprehension experiments need to explore the exact sense of local coherence operative in Hindi and also ascertain whether lacunae in current grammar formalisms in vogue can be rectified using models like trigrams, which incorporate sequence information. Levy (2008) points out that the local sequences like *the player thrown/tossed the frisbee* used in experiments by Tabor and colleagues present difficulties for Surprisal Theory and recommend factoring in uncertainty about previous words into PCFG surprisal estimates. Moreover, the overwhelming success of trigram surprisal merits a more nuanced interpretation. In a recent survey, Staub (2015) summarizes the role of lexical predictability as a graded effect, where instead of discretely predicting a single word, the comprehension system activates a series of potential upcoming words. Moreover, lexical predictability effects occur either at the very early stages of lexical access or pre-lexical stages (processing visual features of letters in the script), rather than at post-lexical stages involving meaning identification. So in future inquiries, we plan to design and deploy a cognitively motivated parsing model which predicts multiple words at every point in time (Section 5.4 discusses the Information Locality Hypothesis, which might be germane to a formal implementation of this idea).

5.3. Case markers

Hindi is a language with rich case marking (a complete list of case markers is provided in Table A.4 of the Appendix). From a cognitive perspective, *prediction* and *interference* are the two types of explanation that have been proposed for effects on account of case markers in language comprehension (Avetisyan et al., 2020; Lewis & Vasishth, 2005). Our survey of the literature indicates that prediction-based explanations of case markers appear to be more plausible than interference-based mechanisms. For Hindi, there is only limited support for similarity-based interference caused by case markers in center-embeddings (Vasishth, 2003). In contrast, predicting the final verb is considered to be a vital part of parsing in head-final languages. In particular, case markers help predict the structure of the upcoming verb phrase (VP) as shown recently in the case of Armenian by Avetisyan et al. (2020). The Hindi ergative marker *ne* predicts the presence of a transitive final verb with perfective marking (Choudhary et al., 2009; Husain et al., 2014). Thus in non-canonical orders where ergative subjects occur after the object, the prediction of VP structure is delayed till the subject is encountered. There is also preliminary evidence that in Hindi, non-canonical ordering of case markers can lead to parsing errors due to local coherence effects (Apurva & Husain, 2019).

5.4. Relationship with information locality

Futrell (2019) pointed out that Surprisal Theory's sole prediction about word order is that less common orders in a language accrue greater processing cost. Futrell modified the theory by positing that the per-word processsing difficulty is proportional to its surprisal given a *lossy memory representation* of the preceding context. This formulation led to the information locality hypothesis (ILH) propounded by Futrell et al. (2020), which articulated novel predictions about word order preferences and online processing. Here we link the ILH predictions made by Futrell et al. (2020) to our findings. In the context of word order preferences, ILH states that going beyond syntactically related words, *all* word pairs with high mutual information tend to be pulled together. Dependency locality is thus subsumed by the more general principle of information locality. Syntactic relatedness is merely a particular instance of informationally-related word pairs being constrained by the grammar of the language. The overwhelming success of our trigram surprisal estimates for the task of reference sentence prediction provides some preliminary evidence for ILH in the case of word order. In our study, trigram surprisal and dependency length make the same prediction in 58% of the cases (*i.e.*, more than random chance). Moreover, only 27.5% of all syntactic dependencies are separated by more than 2 words and hence fall outside of a single trigram window (the rest are thus

[22] We are indebted to Prof. Shravan Vasishth for this idea and references.
[23] We are indebted to Prof. Whitney Tabor for this suggestion.

directly captured by trigrams). Alternatively, it is plausible that symbol grammars are inadequate to a certain degree for capturing dependencies native speakers rely on during processing. Thus future inquiries need to investigate whether particular properties of grammar formalisms can be connected to mutual information.[24]

5.5. Implications for language production

Recent work has pointed to the need for greater synergy between language comprehension and production research, as these processes do not occur in isolation but in fact are tightly coupled (Pickering & Garrod, 2007, 2013). In this context, the Production-Distribution-Comprehension (PDC) account (MacDonald, 1999, 2013) linking language production, comprehension, and typology posits ease of production as the *sole factor* driving language production. Further, PDC conceives of language comprehension as being influenced by the distribution of forms in language obtained as a result of production. Speakers acquire such distributions and subsequently use them to guide language comprehension. Levy and Gibson (2013) state that for PDC to become a computationally viable theory of language processing, it must make *localized* and *incremental* predictions of processing difficulty. They further add that surprisal is an existing measure of comprehension difficulty that can be used to formalize the PDC comprehension component. Thus our surprisal estimates which provide per-word values of lexical and syntactic predictability can be integrated into a computational model of the PDC account in order to test cognitively-grounded hypotheses pertaining to syntactic choice and language production. As per the PDC account, the dominance of expectation measures for Hindi syntactic choice would be a factor pertaining to production ease as opposed to facilitating comprehension for listeners. The close coupling between our syntactic choice results and findings in Hindi comprehension needs to be explored from this standpoint using spontaneous production experiments.

6. Conclusions and future work

In this paper, we show that minimization of dependency length predicts preverbal syntactic choice in written Hindi, even when information status and expectation-based control variables, *viz.*, surprisal estimates (from trigram and incremental dependency parsing models) are present in a logistic regression model for reference sentence prediction. Dependency length is a weak predictor compared to both the surprisal measures as reflected from its regression coefficient as well as accuracy in predicting corpus sentences amidst grammatical variants. This could be because decay, the underlying cognitive construct behind locality, does not have robust empirical evidence supporting it (Engelmann et al., 2019; Oberauer & Lewandowsky, 2013). The alternate explanation of interference proposed in the literature on dependency completion (Engelmann et al., 2019; Jager et al., 2017; Vasishth et al., 2019) should be investigated more thoroughly in future inquiries. Trigram surprisal is strongest predictor in terms of prediction accuracy and the prediction performance of parser surprisal lies between that of dependency length and trigram surprisal. Both trigram surprisal and dependency length tend to be effective when the absolute dependency length difference between reference-variant pairs is high. We connect trigram surprisal effects to local coherence considerations discussed by Tabor et al. (2004). Finally, we show that dependency minimization constraints are factored while choosing references sentences having certain non-canonical word orders (not predicted by surprisal measures due their bias for frequent structures) over variants containing the canonical subject-object order. This result mirrors the results reported by Vasishth (2004) in their study of the comprehension of Hindi object-fronted orders. In such constructions, dependency length is effective even when information status considerations are present in the model. Finally, our analyses revealed that the presence of case markers can potentially override the impact of long dependencies.

As mentioned at the start of this section, currently, we use information status score as a control factor. Arnold et al. (2013) suggest methods by which information status can be connected to cognitive processing. For example, properties like salience and givenness can be connected to predictability, already modelled by surprisal. In fact our dependency parser surprisal estimates currently derived from words and their morpho-syntactic features can potentially be augmented by information status features along the lines of Cahill and Riester (2009) and automatic focus annotations described in Ziai and Meurers (2018). More recently, Engelmann et al. (2019) demonstrated that the cue-based retrieval framework proposed by Lewis and Vasishth (2005) augmented with discourse salience achieved a better fit to reading time data from a wide variety of sentence processing experiments than the original model. Thus more sophisticated surprisal estimates and computational models encoding information status considerations can be deployed to investigate real time processing in constituent ordering.

Recent research has demonstrated that case markers might play a crucial role in language evolution. Tily (2010) showed that Old English (with more SOV-like order) had case markers and higher average dependency length compared to Middle and Modern English. However, in the course of evolution to the SVO order of Modern English, Old English lost case markers and word order freedom concomitant with the drive to minimize dependency length. The diachronic evolution of expectation effects needs to be charted out. In another line of work, Ferrer-i-Cancho presented mathematical proofs to the effect that SOV languages maximize predictability of the final verb (Ferrer-i-Cancho, 2017), while SVO languages tend to minimize online memory load (Ferrer-i-Cancho, 2015).[25] The high impact of dependency length on English constituent ordering choices (Rajkumar et al., 2016) interpreted in conjunction with the current study showing the weak impact of dependency length on Hindi constituent ordering choices lend further empirical support to this idea. Thus, in the more established view, languages might have evolved along the twin axes of predictability maximization and dependency length minimization, with a potential trade-off between the two factors. However, an alternative way of thinking is provided by Futrell (2019) and Futrell et al. (2020) with their information locality hypothesis discussed in Section 5.4, where both these axes can be subsumed by the more general principle of maximizing information locality. This framework could be a more unified way of theorising about word order across different languages, with the variations across them perhaps corresponding to differences in how informational links are distributed. Examining these frameworks further and for a wider variety of languages to obtain a more comprehensive psycholinguistic theory of word order remains a key direction for future research.

Acknowledgements

We acknowledge support from the Faculty Initiation Grant of IISER Bhopal (Project No. INST/HSS/2018096) and extramural funding from the Cognitive Science Research Initiative, Department of Science and Technology, Government of India (DO: DST/CSRI/2018/263). We are grateful to Professors KP Mohanan, Shravan Vasishth, Whitney Tabor, and Silvia Gennari (Cognition Associate Editor) for their invaluable feedback. Thanks to Ayush Jain and Arpit Agrawal for their help in running the dependency parser. Finally, we thank Rupesh Pandey for help with sentence rating.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cognition.2021.104959.

[24] Thanks to Reviewer 2 for pointing this out.

[25] For a critique of this idea, please refer to Alday (2015).

References

Agnihotri, R. K. (2007). *Hindi: An essential grammar. essential grammars*. Oxfordshire: Routledge.

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

Alday, P. M. (2015). Be careful when assuming the obvious. *Language Dynamics and Change, 5*(1), 138–146.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated theory of the mind. *Psychology Review, 111*(4), 1036–1060.

Anderson, J. R., & Paulson, R. (1977). Representation and retention of verbatim information. *Journal of Verbal Learning and Verbal Behavior, 16*(4), 439–451.

Anttila, A., Adams, M., & Speriosu, M. (2010). The role of prosody in the English dative alternation. *Language and Cognitive Processes, 25*(7-9), 946–981.

Apurva, A., & Husain, S. (2019). Local coherence and case-marker exchange cause parsing errors in Hindi. In *The 6th annual conference of cognitive science, BITS Pilani, Goa*. India: Association for Cognitive Sciences.

Arnold, J. E. (2011). Ordering choices in production: For the speaker or for the listener? In E. M. Bender, & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing* (pp. 199–222). Stanford University: CSLI Publishers.

Arnold, J. E., Kaiser, E., Kahn, J. M., & Kim, L. K. (2013). Information structure: Linguistic, cognitive, and processing approaches. *WIREs Cognitive Science, 4*(4), 403–413.

Arnold, J. E., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language, 76*, 28–55.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language, 112*, 104087.

Baayen, R. H. (2008). *Analyzing linguistic data* (1 edition). Cambridge: Cambridge University Press.

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 LREC 2002* (pp. 819–827). Lancaster University.

Bhatia, S., & Husain, S. (2018). Forgetting effects in a head-final language: Evidence from Hindi. In *Proceedings of the 31st annual cuny sentence processing conference*. California, USA: UC Davis.

Bhatia, S., & Husain, S. (2019). Prediction failure and local coherence in a head-final language. In *Proceedings of the psycholinguistics in Iceland - parsing and prediction (PIPP) conference, Rejkavik. Iceland*.

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, ACL-IJCNLP '09* (pp. 186–189). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bicknell, K., Levy, R., & Demberg, V. (2009). Correcting the incorrect: Local coherence effects modeled with prior belief update. *Annual Meeting of the Berkeley Linguistics Society, 35*, 13–24.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition, 21*, 47–67.

Bock, K., & Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 467–484.

Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research, 2*(1), 1–12.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes, 26*(3), 301–349.

Breslow, N., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9–25.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. *Cognitive Foundations of Interpretation*, 69–94.

Butt, M., & King, T. H. (1996). Structural topic and focus without movement. In M. Butt, & T. H. King (Eds.), *Proceedings of the first LFG conference*. Stanford: CSLI Publications.

Cahill, A., & Riester, A. (2009). Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 817–825). Suntec, Singapore: Association for Computational Linguistics.

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language, 89*, 244–254. Speaking and listening: Relationships between language production and comprehension.

Choi, H.-w. (2007). Length and order: A corpus study of Korean dative-accusative construction. *Discourse and Cognition, 14*(3), 207–227.

Choudhary, K. (2011). *Incremental argument interpretation in a split ergative language: Neurophysiological evidence from Hindi. MPI series in human cognitive and brain sciences*. Max Planck Institute for Human Cognitive and Brain Sciences Leipzig: MPI for Human Cognitive and Brain Sciences.

Choudhary, K. K., Schlesewsky, M., Roehm, D., & Bornkessel-Schlesewsky, I. D. (2009). The N400 as a correlate of interpretively relevant linguistic rules: Evidence from Hindi. *Neuropsychologia, 47*(13), 3012–3022.

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

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*. Chicago: University of Chicago Press: Blackwell Oxford.

Corley, S., & Crocker, M. (2000). *The modular statistical hypothesis: Exploring lexical category ambiguity* (pp. 135–160).

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research, 29*(6), 647–669.

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

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.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology, 40*(4), 296–340.

Ferreira, V. S., & Yoshita, H. (2003). Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research, 32*(6), 669–692.

Ferrer-i-Cancho, R. (2015). The placement of the head that minimizes online memory. *Language Dynamics and Change, 5*(1), 114–137.

Ferrer-i-Cancho, R. (2017). The placement of the head that maximizes predictability. an information theoretic approach. *Glottometrics, 39*, 38–71.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing.. In *Proceedings of the 3rd annual workshop on cognitive modeling and computational linguistics* (pp. 61–69).

Futrell, R. (2019). 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.

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.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336–10341.

Gallo, C. G., Jaeger, T. F., & Smyth, R. (2008). Incremental syntactic planning across clauses. In *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 845–850).

Gambhir, V. (1981). *Syntactic restrictions and discourse functions of word order in standard Hindi*. Philadelphia: University of Pennsylvania. PhD thesis.

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.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407.

Gildea, D., & Jaeger, T. F. (2015). *Human languages order information efficiently*. CoRR, abs/1510.02823.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science, 34*(2), 286–310.

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, NAACL '01* (pp. 1–8). Pittsburgh, Pennsylvania: Association for Computational Linguistics.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. New York: Cambridge University Press.

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

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Husain, S., Bhatt, R., & Vasishth, S. (2013). Towards a psycholinguistically motivated dependency grammar for Hindi. In *Proceedings of the Second International Conference on Dependency Linguistics (DepL ing 2013)* (pp. 108–117). Prague, Czech Republic: Charles University in Prague, Matfyzpress, Prague, Czech Republic.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLOS ONE, 9*(7), 1–14.

Husain, S., Vasishth, S., & Srinivasan, N. (2015). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research, 8*(2.), 1–12.

Jaeger, T. F. (2011). Corpus-based research on language production: Information density and reducible subject relatives. In E. M. Bender, & J. E. Arnold (Eds.), *Language from a cognitive perspective: grammar, usage, and processing* (pp. 161–197). Stanford: CSLI Publishers.

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

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

Jager, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language, 94*, 316–339.

Jager, L. A., Mertzen, 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.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '02* (pp. 133–142). New York, NY, USA: ACM.

Kachru, Y. (1982). Conjunct verbs in Hindi-Urdu and persian. *South Asian Review, 6*(3), 117–126.

Kachru, Y. (2006). *Hindi. London oriental and African language library*. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Kempen, G., & Harbusch, K. (2008). Comparing linguistice judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in german clauses. In A. Streube (Ed.), *The discourse potential of underspecified structures* (pp. 179–192). Berlin: Walter de Gruyter.

Kidwai, A. (2000). *XP-adjunction in universal grammar: Scrambling and binding in Hindi-Urdu: Scrambling and Binding in Hindi-Urdu. oxford studies in comparative syntax*. Oxford: Oxford University Press.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguists Research, 29*(6), 627–645.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MA: MIT Press.

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

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence processing, current issues in the psychology of language* (1 edition). UK: Psychology Press.

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of russian relative clauses. *Journal of Memory and Language, 69*(4), 461–495.

Levy, R., & Gibson, E. (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology, 4*, 229..

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language, 68*(2), 199–222.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 1–45.

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*(10), 447–454.

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science, 40*(6), 1382–1411.

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

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews, 21*, 171–193.

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 177–196). NJ, USA: Lawrence Erlbaum Associates Publishers.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology, 4*(226), 1–16. Published with commentaries in Frontiers.

Mahajan, A. K. (1990). *The a/a-bar distinction and movement theory*. MA: Massachusetts Institute of Technology. PhD thesis.

Mishra, R. K., Pandey, A., & Srinivasan, N. (2011). Revisiting the scrambling complexity hypothesis in sentence processing: A self-paced reading study on anomaly detection and scrambling in Hindi. *Reading and Writing, 24*(6), 709–727.

Mohanan, K., & Mohanan, 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* (pp. 153–184). Stanford, CA: Center for the Study of Language and Information.

Mohanan, T. (1994). Case ocp: A constraint on word order in Hindi. In M. Butt, T. H. King, & G. Ramchand (Eds.), *Theoretical perspectives on word order in South Asian languages* (pp. 185–216). Stanford, CA: Center for the Study of Language and Information.

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using bayesian hierarchical modeling. *Journal of Memory and Language, 99*, 1–34.

Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. *Comput. Linguist., 34*(4), 513–553.

Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General, 142*(2), 380–411.

Patil, U., Kentner, G., Gollrad, A., Kügler, F., Féry, C., & Vasishth, S. (2008). Focus, word order and intonation in Hindi. *Journal of South Asian Linguistics, 1*(1), 55–72.

Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44* (pp. 433–440). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105–110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*, 329–347.

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language, 42*(2), 168–182.

Qian, T., & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive Science, 36*(7), 1312–1336.

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.

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

Ranjan, S., Agarwal, S., & Rajkumar, R. (2019). 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.

Reali, F., & Christiansen, M. (2007). Word chunk frequencies affect the processing of pronomial object-relative clauses. *Quarterly journal of experimental psychology (2006), 60*, 161–170.

Ros, I., Santesteban, M., Fukumura, K., & Laka, I. (2015). Aiming at shorter dependencies: The role of agreement morphology. *Language, Cognition and Neuroscience, 30*(9), 1156–1174.

Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in persian separable complex predicates: Evidence for expectation and memory-based accounts. *Frontiers in Psychology, 7*, 403.

Shain, C., van Schijndel, M., Gibson, E., & Schuler, W. (2016). Exploring memory and processing through a gold standard annotation of Dundee. In *Proceedings of CUNY 2016*. Gainesville, Florida, USA: University of Florida.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal, 27*, 379–423.

Snider, N. (2009). Similarity and structural priming. In *Proceedings of the 31th annual conference of the cognitive science society* (pp. 815–820).

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-np shift. *Journal of Memory and Language, 39*(3), 392–417.

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

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proc. ICSLP-02*.

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory, 1*, 113–150.

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.

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

Temperley, D., & Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Reviews of Linguistics, 4*, 67–80.

Tily, H. (2010). *The role of processing complexity in word order variation and change, PhD thesis*. Stanford University. unpublished thesis.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*(2), 407–430.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language, 55*(2), 157–166.

van Schijndel, M., Nguyen, L., & Schuler, W. (2013). An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proceedings of CMCL 2013, Sofia, Bulgaria*. Association for Computational Linguistics.

van Schijndel, M., & Schuler, W. (2013). An analysis of frequency- and memory-based processing costs. In *Proceedings of NAACL-HLT 2012*. Atlanta, Georgia, USA: Association for Computational Linguistics.

Vasishth, S. (2003). *Working memory in sentence comprehension: Processing Hindi center embeddings. Outstanding dissertations in linguistics*. Oxfordshire: Taylor & Francis.

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

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language, 82*(4), 767–794.

Vasishth, S., Mertzen, D., Jaeger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151–175.

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences, 23*(11), 968–982.

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.

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.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition, 85*, 79–112.

Wasow, T. (2002). *Postverbal behavior*. Stanford: CSLI Publications.

Yadav, H., Vaidya, A., & Husain, S. (2017). *Keeping it simple: Generating phrase structure trees from a Hindi dependency treebank* (pp. 123–133). Citeseer, NJ, USA: TLT.

Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition, 81*, B45–B55.

Ziai, R., & Meurers, D. (2018). Automatic focus annotation: Bringing formal pragmatics alive in analyzing the information structure of authentic data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 117–128). New Orleans, Louisiana: Association for Computational Linguistics.

Appendix A: Supplementary Information and Figures

Dependency Length Calculation

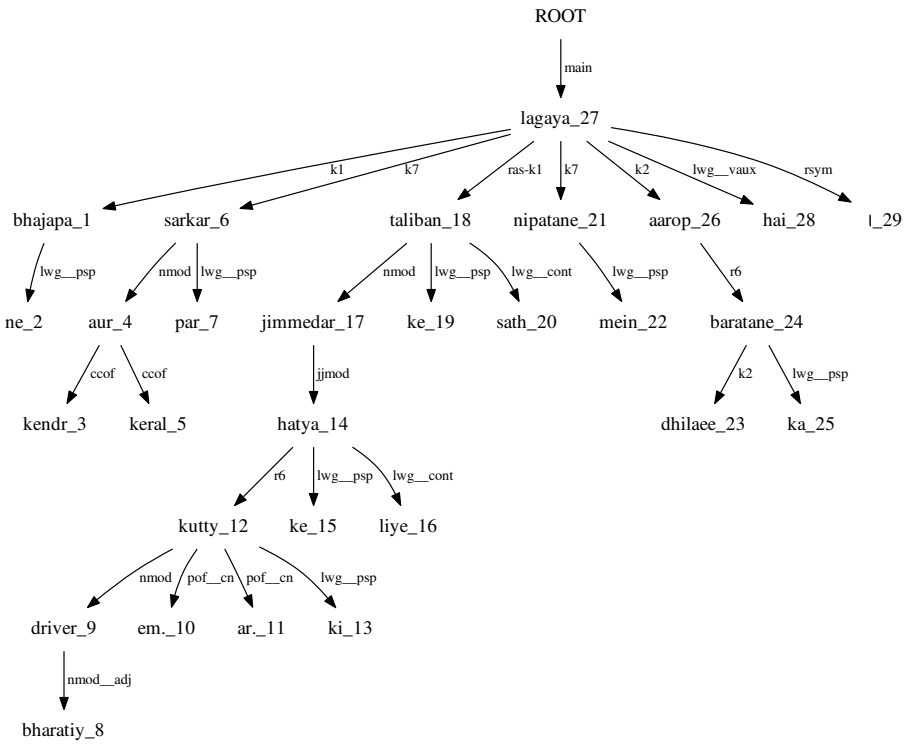


Figure A.1: Example dependency tree (word_position) corresponding to Example 9

Word Index	Word	Head Index	Dependency Length (|Head Index - Word Index| - 1)
1	bhajapa	27	25
2	ne	1	0
3	kendr	4	0
4	aur	6	1
5	keral	4	0
6	sarkar	27	20
7	par	6	0
8	bharatiy	9	0
9	driver	12	2
10	em	12	1
11	ar	12	0
12	kutty	14	1
13	ki	12	0
14	hatya	17	2
15	ke	14	0
16	liye	14	1
17	jimmedar	18	0
18	taliban	27	8
19	ke	18	0
20	sath	18	1
21	nipatane	27	5
22	mein	21	0
23	dhilaee	24	0
24	baratane	26	1
25	ka	24	0
26	aarop	27	0
27	lagaya	ROOT	–
28	hai	27	0
Total			**68**

Table A.1: Dependency length calculation for tree in Figure A.1 (corresponding to Example 9)

(9) [bhajapa=ne] [*kendr aur keral sarkar-par*] [bharatiy driver em.ar. kutty-ki
BJP=ERG centre and Kerala government-on Indian driver M.R. Kutty-GEN
hatya=ke liye jimmedar taliban=ke sath] [nipatane=mein] [dhilaee baratane=ka
murder=GEN for responsible Taliban=GEN with deal laxness action=GEN
aarop] lag-aa-ya hai
accusation do-CAUS-PFV.M be.PRS.SG

BJP has accused the centre and Kerala governments over laxness in dealing with Taliban, accused in the murder of Indian driver M. R. Kutty.

Constructions

Table A.2 depicts some of the main constructions in our dataset. Table A.3 shows examples of reference sentences, their variants as well as variants which got excluded as a result of filtering

Construction	Active	Passive
	6733/143721	853/15170
Subject present	5989/132605	116/2781
No subject	744/1116	737/12389
Direct object	1909/66867	552/11218
No direct object	4824/76854	301/3952
Indirect object	660/21737	109/2884
No indirect object	6073/121984	744/12286
S-DO-V order	1454/33808	27/767
Direct object fronted	127/27181	6/584
S-IO-V order	475/10774	6/413
Indirect object fronted	97/9227	5/252
Subject Relative Clause	40/378	1/1
Object Relative Clause	17/119	4/14
Overt complementizer	1394/9840	28/120
No complementizer	5339/133881	825/15050
Conjunct verbs (CV)	4075/104351	531/11669
CV-Adjacent	3647/893773	470/3329
CV-intervening	428/14978	61/8340

Table A.2: Selected constructions in our dataset (#reference sentences/#variants)

using dependency relations.

Locality and Non-Locality Cases

The following sentences illustrate the success of parser surprisal in non-locality and zero-locality cases for the same reference sentence involving the conjunct verb *ghoshna ki* discussed in the previous section. However, unlike the variant depicted in Example 6b, here we focus on remaining five variants where both parts of the conjunct verb are together (dependency length, trigram, surprisal, and parser surprisal indicated alongside):

(10) a. [brihaspativar=ko] [company=ki vaarshik aam baithak=me] [reliance=ke adhyaksh Thursday=at company-GEN annual general meeting=LOC reliance-GEN chairman mukesh ambani=ne] is baat=ki **ghoshna** **k-i** (40, 38.65, 5.19)
Mukesh Ambani=ERG this matter=GEN announcement do-PFV.F

On Thursday, the chairman of Reliance, Mukesh Ambani, announced this matter at the company's annual general meeting.

b. company ki vaarshik aam baithak mein brihaspativar ko reliance ke adhyaksh mukesh ambani ne is baat ki **ghoshna ki** (36, 37.09, 5.25)

c. reliance ke adhyaksh mukesh ambani ne brihaspativar ko company ki vaarshik aam baithak mein is baat ki **ghoshna ki** (32, 36.45, 5.25)

d. reliance ke adhyaksh mukesh ambani ne company ki vaarshik aam baithak mein brihaspativar ko is baat ki **ghoshna ki** (32, 36.67, 5.20)

e. company ki vaarshik aam baithak mein reliance ke adhyaksh mukesh ambani ne brihaspativar ko is baat ki **ghoshna ki** (36, 37.06, 5.16)

f. brihaspativar ko reliance ke adhyaksh mukesh ambani ne company ki vaarshik aam baithak mein is baat ki **ghoshna ki** (40, 36.68, 4.77)

Type	Sentence
Reference	*vaise Zulfikaar ka parivaar khud ko Bhaaratiya nagarik bataataa rahaa* Anyway Zulfikaar-GEN. family self-REF Indian citizen claim-IMFV.m.sg Anyway Zulfikaar's family claimed themselves to be Indian citizens.
Variants	vaise khud ko Zulfikaar ka parivaar Bhaaratiya nagarik bataataa rahaa khud ko vaise Zulfikaar ka parivaar Bhaaratiya nagarik bataataa rahaa
Excluded	Bhaaratiya nagarik khud ko Zulfikaar ka parivaar vaise bataataa rahaa Bhaaratiya nagarik vaise khud ko Zulfikaar ka parivaar bataataa rahaa
Reference	*raahat shiviron me math ke log bhojan uplabdh karaa rahe hain* Relief camp-LOC monastery-GEN people food arrange do-IMFV/pl.pst People in the monastery arranged food for the relief camps.
Variants	raahat shiviron me bhojan math ke log uplabdh karaa rahe hain math ke log raahat shiviron me bhojan uplabdh karaa rahe hain
Excluded	uplabdh bhojan raahat shiviron me math ke log karaa rahe hain uplabdh bhojan math ke log raahat shiviron me karaa rahe hain
Reference	*vahaan se sabhi vidhaayak train ke zariye Bhuvaneshwar pahuche* There-INS all MLA-pl train INS Bhuvaneshwar reach-PFV.m/f/pl All the legislators reached Bhubaneshwar by train from there.
Variants	vahaan se train ke zariye sabhi vidhaayak Bhuvaneshwar pahuche sabhi vidhaayak vahaan se train ke zariye Bhuvaneshwar pahuche
Excluded	Bhuvaneshwar train ke zariye vahaan se sabhi vidhaayak pahuche Bhuvaneshwar train ke zariye sabhi vidhaayak vahaan se pahuche

Table A.3: Example reference sentences (italicized) and their variants (correct and excluded)

For the reference-variant pairs above, parser surprisal predicts the reference sentence (Example 10a) over each of the first three variants correctly (Examples 10b-10d), whereas trigram surprisal and dependency length pick the variant in all the pairs above. However, due to competition from a stronger trigram surprisal predictor, the efficacy of parser surprisal on its own does not translate into the model successfully choosing the reference sentence over any of these variants.

Hindi Case Markers

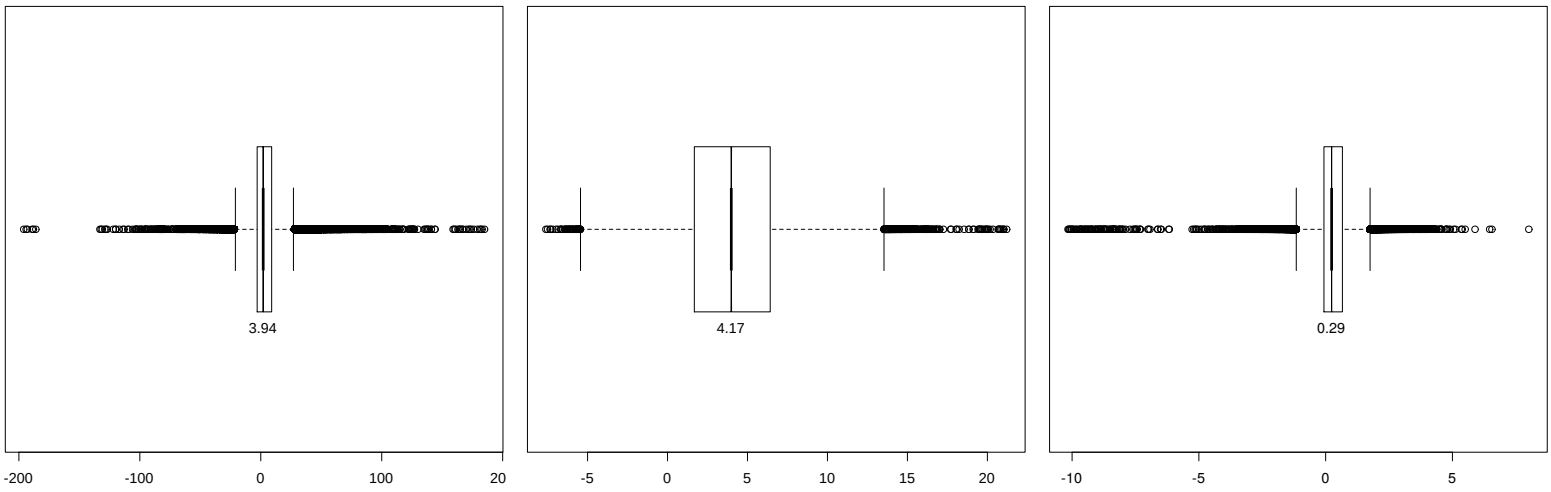
Marker	**Case** (Gloss)	**Grammatical Function**
ϕ	nominative (NOM)	subject/object
ne	ergative (ERG)	subject
ko	accusative (ACC)	object
	dative (DAT)	subject/indirect object
se	instrumental (INS)	subject/oblique/adjunct
ka/ki/ke	genitive (GEN)	subject (infinitives) specifier
mẽ/par/tak	locative (LOC)	oblique/adjunct

Table A.4: Hindi case markers (Butt and King 1996)

Regression Results

Normalized predictor values	Entire data (158891)	Conjunct verbs (116020)	DO-fronted (1663)	IO-fronted (1353)
0 0 0	0.505 (0.499, 0.509)	0.506 (0.499, 0.513)	0.505 (0.470, 0.541)	0.541 (0.493, 0.588)
1 0 0	0.005 (0.004, 0.005)	0.003 (0.003, 0.004)	0.062 (0.046, 0.085)	0.006 (0.004, 0.012)
0 1 0	0.348 (0.341, 0.356)	0.374 (0.365, 0.383)	0.389 (0.333, 0.447)	0.377 (0.300, 0.462)
0 0 1	0.434 (0.425, 0.442)	0.425 (0.415, 0.435)	0.369 (0.320, 0.421)	0.515 (0.442, 0.589)

Table A.5: Probability of predicting the positive class using a model trained using normalized predictor vectors (95% CI given in parenthesesin each cell; # data points given in the header row)



(a) Dependency length (b) trigram surprisal (c) Parser surprisal

Figure A.2: Mean of difference between mean predictor values of variants and reference sentences (95% classification intervals indicated)