Uniform Information Density Effects on Syntactic Choice in Hindi

Ayush Jain USC
Vishal Singh NYU
Sidharth Ranjan IIT Delhi
ayushj240,vishal.singh5846,sidharth.ranjan03@gmail.com

Rajakrishnan Rajkumar IISER Bhopal
rajak@iiserb.ac.in
Sumeet Agarwal IIT Delhi
sumeet@iitd.ac.in

Abstract

According to the UNIFORM INFORMATION DENSITY (UID) hypothesis (Levy and Jaeger, 2007; Jaeger, 2010), speakers tend to distribute information density across the signal uniformly while producing language. The prior works cited above studied syntactic reduction in language production at *particular choice points* in a sentence. In contrast, we use a variant of the above UID hypothesis in order to investigate the extent to which word order choices in Hindi are influenced by the drive to minimize the variance of information across *entire sentences*. To this end, we propose multiple lexical and syntactic measures (at both word and constituent levels) to capture the uniform spread of information across a sentence. Subsequently, we incorporate these measures in machine learning models aimed to distinguish between a naturally occurring corpus sentence and its grammatical variants (expressing the same idea). Our results indicate that our UID measures are not a significant factor in predicting the corpus sentence in the presence of lexical surprisal, a competing control predictor. Finally, in the light of other recent works, we conclude with a discussion of reasons for UID not being suitable for a theory of word order.

1 Introduction

The Uniform Information Density (henceforth UID) hypothesis states that language production exhibits a preference for distributing information uniformly across a linguistic signal. This hypothesis has a long history in the literature and Ferrer-i-Cancho (2017) traces the idea to the pioneering work of August and Gertraud Fenk (Fenk and Fenk-Oczlon, 1980) and developed further in subsequent articles (Fenk-Oczlon, 2001, for an overview). In recent years, this hypothesis has gained substantial traction with the work on syntactic reduction done by Florian Jaeger and colleagues (Levy and Jaeger, 2007; Jaeger, 2010). They show that speakers achieve uniformity of information across utterances either by omitting optional function words (like the *that* complementizer) or by explicitly mentioning them. In contrast to the two prior works cited above, which look at information density at *particular choice points* in language production, we examine a variant of the UID hypothesis stated above in the case of *entire sentences* created by syntactic alternations.

In this work, we test the hypothesis that reference sentences obtained from a corpus of naturally occurring written text exhibit greater uniformity in the spread of information in comparison to grammatical variants expressing the same idea. To this end, inspired from Collins (2014), we propose five distinct UID measures quantifying the uniformity of information density at both syntactic and lexical levels. We test two different versions of these measures at word as well as constituent boundaries. We examine the impact our UID measures in predicting syntactic choice in Hindi, an Indo-Aryan language with predominantly SOV word order and case-marking postpositions. This is the first work on the Hindi language (to the best of our knowledge), which studies its information-theoretic properties pertaining to syntac-

The first three authors listed are joint first authors. Ayush Jain and Vishal Singh undertook this project while they were undergraduate students at IIT Delhi.

Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing, pages 38–48
Santa Fe, New Mexico, USA, August 25, 2018.

tic choice. In comparison to English (SVO order and prepositions), Hindi has relatively flexible word order (Agnihotri, 2007; Kachru, 2006).

Our study uses written data from the Hindi-Urdu Treebank (HUTB) corpus (Bhatt et al., 2009) consisting of newswire text. Hence the sentences used in our study are by default set in a given context. In addition to production ease, the language production system also factors in communicative considerations pertaining to facilitating comprehension for listeners (i.e. *audience design*) and for the speakers themselves (Jaeger and Buz, in press). Moreover, written text is often edited, taking into account comprehensibility considerations explicitly[1]. From the perspective of online language comprehension, processing difficulty is quantified by surprisal (Hale, 2001; Levy, 2008). We examine whether the UID measures we defined are significant predictors of syntactic choice even amidst lexical and syntactic surprisal as control factors (modelling comprehension considerations). Our experiments primarily involved the task of classifying Hindi data into reference sentences and artificial generated variants created by linearizing dependency trees corresponding to reference sentences in the HUTB corpus. Our UID measures were deployed as features in machine learning models to perform this binary classification task.

Our results indicate that logistic regression models containing lexical surprisal along with our lexical and syntactic UID measures (across words as well as constituents) do not significantly outperform a strong baseline model containing only lexical surprisal (estimated using a simple trigram model over words). Weak effects of both lexical and syntactic UID measures are attested in some non-canonical word order sequences involving object fronting. However, these are not in the expected direction *i.e.*, corpus sentences are characterized by spikes and troughs in information across words compared to their artificially generated variants. This result is very similar to that reported in the work of (Maurits et al., 2010), where the authors showed that object-first orders are in conflict with their formulation of the UID hypothesis. Using a corpus study as well as results from judgement tasks, they show that such orders cause troughs in the signal compared to other orders because of the disproportionate amount of information clustered around the object, making subsequent elements of the sentence redundant. They also point out the failure of their version of the UID hypothesis in the case of SOV languages. They attribute it to the presence of other stronger factors in such languages. On a related note, Ferrer-i-Cancho (2017) discuss how predicting the final verb is a stronger processing pressure in verb-final languages compared to other competing principles like dependency length minimization. Our result demonstrating lexical surprisal as a robust predictor of Hindi syntactic choice, adds support to predictability as a strong determinant of syntactic choice. Thus we conclude that the UID hypothesis (as defined by our measures) does not shape word order choices in Hindi when other control factors like predictability are considered. We discuss possible reasons for this by alluding to the work of (Ferrer-i-Cancho, 2017). This recent work suggests that UID might not be appropriate for a theory of word order of languages and UID might be restricted to account for syntactic reduction phenomena only.

The paper is structured as follows. Section 2 offers a brief background on the UID hypothesis and surprisal. Section 3 describes the UID measures we proposed as part of this work. Section 4 provides details of the datasets and models we used for testing our hypotheses. Section 5 presents the experiments conducted as part of the study and Section 6 discusses the implications of the results obtained for a theory of word order. Finally, Section 7 summarizes the conclusions as well as reflects on possible directions of future inquiry.

2 Background

The UNIFORM INFORMATION DENSITY principle discussed by (Jaeger, 2010) predicts that language production is optimized to distribute information uniformly across the utterance without exceeding the capacity of the communication channel. Claude Shannon's definition of information (Shannon, 1948) is adopted in this work. Information is defined as the negative log of the conditional probability of the linguistic unit (usually a word) in a given context. In context of omission or mention of the optional *that*-complementizer in English, Jaeger hypothesized that if the information density at the beginning

[1]In early Natural Language Generation research, editing performed by authors was considered to be akin to the self-monitoring component in Willem Levelt's 1989 model of human language production (Neumann and van Noord, 1992).

of a complement clause (CC) is high enough to exceed the capacity of the communication channel, then native speakers tend to explicity mention the *that*-complementizer at the start of the complement clause. The reason for this is the impact of the high frequency word *that* in reducing the information density at the CC onset. Conversely, for a CC with low information density at the beginning, omitting the *that*-complementizer would achieve the effect of increasing the information density at this choice point. Jaeger tested this hypothesis by examining *that*-reduction in the Switchboard corpus of English conversational speech. This study conclusively showed that information density is a significant predictor of *that*-mention (or omission) even while controls based on competing hypotheses were included in the statistical model to predict complementizer choice in spoken English.

Surprisal is mathematically equivalent to information density defined for language production, but it is an indicator of human sentence comprehension load based on different theoretical assumptions about activation allocation (Hale, 2001; Levy, 2008). We use two standard definitions of surprisal in this work as described below:

1. **Lexical surprisal** for word $k+1$ is defined using the conditional probability of a word given its two word sentential context and estimated using a simple trigram model over words. Mathematically, surprisal of the $(k+1)^{th}$ word, w, $S_{k+1} = -\log P(w_{k+1}|w_{k-1}, w_k)$.

2. **Syntactic surprisal** is computed using the probabilistic incremental dependency parser developed by (Agrawal et al., 2017), which is 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 most probable parses at each word as it proceeds through the sentence. A maximum-entropy classifier 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. The syntactic surprisal of the $(k+1)^{th}$ word is computed as the log-ratio of the sum of probabilities of maintained parser states at word k to the same sum at word $k+1$.

3 UID Measures

This section describes in detail the five distinct UID measures (two normalized and three unnormalized) we propose as part of this work, in accordance to our version of the UID hypothesis pertaining to entire sentences (as opposed to particular choice points in Jaeger's work). The unnormalized measures are along the lines of UID measures proposed in (Collins, 2014) and their normalized counterparts are our own original contribution. In our work, contextual probabilities used to quantify information density were estimated using lexical as well as syntactic surprisal models described in the previous section. Notation: N is the number of words in a sentence, id_i is the information density (negative lexical/syntactic log-prob) of the i^{th} word of the sentence and μ is defined as the mean information density of the sentence, i.e., $\mu \equiv \frac{1}{N}\sum_{i=1}^{N} id_i$.

1. **Global UID Measure**: $UIDglob = -\frac{1}{N}\sum_{i=1}^{N}(id_i - \mu)^2$

   This measure encapsulates the negative **variance** of information present in a sentence. This is the crux of UID hypothesis which states that the information content at different points in a sentence should not vary much. Thus negative variance appears to be the most straightforward way to capture the uniformity in information density in the sentence.

2. **Local UID Measure**: $UIDloc = -\frac{1}{N}\sum_{i=2}^{N}(id_i - id_{i-1})^2$

   This score represents the negative mean-squared increase or decrease of information content per word, relative to the preceding word. This measure looks at the local uniformity in information in comparison to $UIDglob$ which looks at the global uniformity of the sentence.

3. **Normalized Global UID Measure**: $UIDglobNorm = -\frac{1}{N}\sum_{i=1}^{N}(\frac{id_i}{\mu} - 1)^2$

   It seems natural to judge the extent of variance in the information density as a fraction of the mean value for a given sentence, rather than in absolute terms. So we normalize the UID measure by the mean of the n-gram information density over all the words in the sentence (μ), to get a measure of (negative) variance relative to the mean.

Predictor(s)	Word-based UID measures				Constituent-based UID measures			
	Lexical		*Syntactic*		*Lexical*		*Syntactic*	
	Weight(s)	%Acc	Weight(s)	%Acc	Weight(s)	%Acc	Weight(s)	%Acc
UIDglob	1.08	72.19	0.40	52.43	-0.88	65.54	-0.02	51.61
UIDloc	0.89	71.22	0.02	49.94	-0.6	53.83	0.08	50.71
UIDglobNorm	-13.11	73.05	-0.09	53.16	-0.81	80.06	0.23	52.81
UIDlocNorm	-2.34	62.38	-0.15	53.9	-0.81	69.76	0.11	53.87
UIDlocPrevNorm	0.00	51.23	0.00	53.58	0.005	39.4	0.00	51.87
Surprisal	-0.81	89.96	-0.11	56.48	-0.81	89.95	-0.11	56.38
Lexical surprisal+UIDglob	-1.00, -0.42	89.99	-0.81, 0.01	89.96	-0.79, -0.18	90.08	-0.74, 0.00	89.96
Lexical surprisal+UIDloc	-0.97, -0.11	90.01	-0.95, 0.04	89.97	-0.80, -0.04	90.00	-0.98, 0.07	90.01
Lexical surprisal+UIDglobNorm	-0.96, -2.18	89.98	-0.81, -0.01	89.96	-0.91, -3.75	90.12	-0.93, 0.13	89.99
Lexical surprisal+UIDlocNorm	-0.98, -0.68	89.99	-0.81, -0.02	89.95	-0.96, -0.50	90.00	-0.74, 0.05	89.98

Table 1: Classification performance of various word and constituent-based UID measures

4. **Normalized Local UID Measure**: $UIDlocNorm = -\frac{1}{N}\frac{\sum_{i=2}^{N}(id_i - id_{i-1})^2}{\mu^2}$

   This measure similarly normalises $UIDloc$ using the mean information density of the sentence.

5. **Previous Word Normalized Local UID Measure**: $UIDlocPrevNorm = -\frac{1}{N}\sum_{i=2}^{N}(\frac{id_i}{id_{i-1}} - 1)^2$

   Here the normalisation is local as well: with respect to the information density of just the preceding word, rather than the mean for the complete sentence. $UIDlocPrevNorm$ is essentially the negation of the mean-squared fractional deviation in information as one traverses the sentence from one word to the next.

4 Data and Models

This section describes the datasets and models we used to test our hypotheses on Hindi. For this study, a total of 8736 labelled, projective dependency trees from the Hindi-Urdu Treebank (HUTB) corpus of written Hindi (Bhatt et al., 2009) were used in our experiments. Variants were generated for each of these trees by randomly permuting preverbal constituents (in the preverbal domain itself). A set of non-corpus variants was creating by randomly choosing utmost 99 such variants corresponding to each HUTB reference sentence. Subsequently, from this set of variants, we filtered out variants containing preverbal dependency relation sequences not attested in the HUTB. This was done as a mechanism to automatically ensure that very unacceptable variants were eliminated from our study. We would like to note that this filtering is not crucial to our results in any way. An earlier unfiltered dataset consisting of all variants also showed similar trends in the results and conclusions.

In total, our dataset consisted of 8736 reference sentences and 175801 variants. We estimated lexical surprisal using trigram models trained on 1 million Hindi sentences from EMILLE Corpus (Baker et al., 2002) using the SRILM toolkit (Stolcke, 2002). Good-Turing discounting was used for smoothing. Syntactic surprisal was estimated using an incremental dependency parser (Agrawal et al., 2017) having state-of-the-art unlabelled dependency parsing accuracy. As discussed in the cited work, the per-word syntactic surprisal estimates were also significant predictors of various measures of reading time.

5 Experiments

In this section, we describe our experiments quantifying the impact of the UID measures (proposed in Section 3) on word order choice.

5.1 Pairwise Classification using Logistic Regression

In order to investigate the individual and collective impact of our UID predictors and controls (lexical and syntactic surprisal), we trained and tested logistic regression models for the binary classification task of choosing corpus sentences vs. non-corpus variants. Since our data set is hugely unbalanced, with many more non-corpus than corpus variants, we use a technique from (Joachims, 2002) to effectively convert it into a balanced setting. We created equal numbers of ordered pairs of the types <*corpus, non-corpus*> and <*non-corpus, corpus*> (both sentences in each pair being variants of each other). Feature values of the first sentence in each ordered pair were subtracted from the second sentence in that pair. For a more detailed illustration, please refer to (Rajkumar et al., 2016). This technique also enables feature values of sentences of differing lengths to be centered. The binary classification task is then to identify

	UIDglob	UIDloc	UIDglobNorm	UIDlocNorm	UIDlocPrevNorm
Lexical surprisal	-0.64	-0.58	0.61	0.35	0.02
Syntactic surprisal	-0.46	-0.40	0.19	0.13	0.01

Table 2: Pearson correlation coefficient between: 1. Lexical surprisal and lexical UID measures (Row 1) 2. Syntactic surprisal and syntactic UID measures (Row 2)

each given pair's type, i.e., given such a pair, identify whether the corpus sentence is the first one or the second one. So this can be seen as a way of training a logistic regression model to do pairwise ranking of sentences. The transformed version of the dataset consisted of 175801 data points. Subsequently, we used the python *scikit-learn* toolkit (v0.16.1) to train logistic regression models on this dataset in order to predict the corpus choice sentence. We performed 27-fold cross-validation for classification, wherein the dataset was divided into 27 distinct parts and each part was tested using models trained on the other 26 sections (100 training iterations using *lbfgs* solver).

Table 1 shows the classification results for models trained on different subsets of our features, including both the lexical and syntactic versions of each feature (at both word and constituent levels). Now, we describe the performance of the word-based lexical and syntactic UID measures (middle column of Table 1). The individual classification results show that the best performing feature is lexical surprisal, which predicts the reference sentence in 89.96% of the cases. The negative sign associated with the regression coefficients of both lexical and syntactic surprisal shows that reference sentences are associated with lower surprisal (lower processing difficulty) compared to the variants. For the UID hypothesis to hold true, the regression coefficients of our UID measures should be associated with a positive sign, signifying greater increase in uniformity of information across the sentence. Now we turn to a discussion of the performance of our UID measures, individually as well as in conjunction with lexical surprisal.

Amongst the lexical UID measures, the normalized global UID measure (UIDglobNorm) is the top performing feature (73.04% classification accuracy), while the raw version (UIDglob) comes very close (72.19% accuracy). The accuracy and direction of the UID measures can be attributed to the correlation of these UID measures with surprisal. Table 2 depicts the Pearson's coefficient of correlation between UID measures at the sentence-level and the corresponding surprisal values. For both lexical and syntactic UID measures, normalization results in the direction of the correlation with surprisal being reversed. Both UIDglob and UIDglobNorm measures are moderately correlated with lexical surprisal and hence their performance is much above random chance. UIDglob has a positive regression coefficient, which shows that reference sentences display tendency to maximize uniformity in the spread of information (i.e. minimize negative variance) compared to variant sentences. This is consistent with the UID hypothesis. UIDglob is negatively correlated with lexical surprisal and hence the direction of the effect is also opposite to that of lexical surprisal, which has a negative coefficient as stated above. However, UIDglobNorm has a negative regression coefficient and this goes counter to the UID hypothesis. Thus, normalization has resulted in a measure which exhibits positive correlation with lexical surprisal, resulting in a tendency to mirror lexical surprisal for the task of discriminating between corpus and non-corpus variants. The raw local UID measure (UIDloc) comes very close with 71.22% performance. Both its normalized counterparts (UIDlocNorm and UIDlocPrevNorm) result in considerably lower performance compared to the raw local measure. This difference can again be explained by normalization resulting in UIDlocNorm having low correlation with lexical surprisal and UIDlocPrevNorm being uncorrelated with lexical surprisal. In fact, previous word-based local normalization (UIDlocPrevNorm) resulted in accuracy close to random chance.

The classification performance of syntactic surprisal is very low (56.48%) compared to that of lexical surprisal. We attribute this is to the fact that our syntactic surprisal estimates are derived from an incremental dependency parser (Agrawal et al., 2017), while the task involves constituent ordering. Consequently, all the syntactic UID measures also result in classification accuracy close to 50%. The direction of the individual syntactic UID measures also mirror the direction of correlation between these UID measures and syntactic surprisal (as in the case of the lexical UID measures).

Now, we turn to interpreting the impact of UID measures in combination with lexical surprisal. In

order to discern the impact of UID measures over and above lexical surprisal (a strong predictor of Hindi syntactic choice), we added each UID measure into a classification model containing only lexical surprisal. The results are shown in the bottom row of Table 1. The differences in classification performance between each UID measure and lexical surprisal is not statistically significant. It is evident from the classification results that all the UID measures (both syntactic and lexical) are not adding anything useful beyond overall lexical surprisal estimated using trigrams. Our results involving global UID measures are in line with similar findings obtained by other researchers for a variety of languages. Gildea and Jaeger (2015) document that for American English (written and spoken), German, Arabic (Modern Standard), Czech and Mandarin Chinese, there is no evidence that the variance of Shannon information across words within sentences is lower than expected by chance.

Another puzzle which emerged out of our experimental results is that the effect of many our UID measures is not in the expected direction. The negative regression coefficients associated with all the lexical UID measures (and two of our syntactic UID measures) in conjunction with lexical surprisal show that the reference sentences actually display a lack of uniformity of information, going counter to the UID hypothesis. In the following section, we present evidence that these quirky effects are linked with structures involving non-canonical word order patterns in Hindi.

5.2 UID and Non-canonical Word Order Patterns

Construction (#data points)	**Predictor(s)**	**Weight(s)**	**%Accuracy**
DO fronting (1741)	Lexical surprisal	-0.52	79.15
	+UIDloc (*lex*)	-0.66, -0.35	80.07
	+UIDloc (*syn*)	-0.67, -0.45	81.05
IO fronting (1460)	Lexical surprisal	-0.14	86.57
	+UIDlocNorm (*lex*)	-0.89, -1.97	87.34
	+UIDlocNorm (*syn*)	-0.88, -1.50	87.05

Table 3: UID and non-canonical word order choices ('+' stands for 'Lexical surprisal +')

Free word order languages are also characterized by non-canonical word order patterns. Hindi largely follows the Subject, Indirect Object (IO), Direct Object (DO) and Verb order (Mohanan and Mohanan, 1994). But both direct and indirect object fronting (involving movement of objects to precede subjects), occur rarely, resulting in marked structures. Vasishth (2004) shows how increased reading times at the verb are attested for Hindi object-fronted structures (compared to the base word order), both with and without context. In the light of this finding, we examine the impact of our word-based UID measures on sentence pairs where the reference sentence has the following non-canonical orders and the variant has the corresponding canonical order: 1. Direct object (DO) fronting 2. Indirect object (IO) fronting.

Table 3 presents our classification results for each construction above for models trained and tested only on data points belonging to those constructions. This was motivated by the plan to examine the properties of these constructions in question. We provide the percentage accuracy and direction of the best-performing UID measure relative to lexical surprisal. In the case of direct object fronting, the UIDloc measures (both lexical and syntactic) outperform all the other UID measures. For indirect object fronting, the normalized local UID measures (both lexical and syntactic) help induce improvements in classification accuracy over lexical surprisal. All the aforementioned accuracy gains over lexical surprisal are statistically significant as per McNemar's χ-square test (two-tailed $p < 0.001$). As evinced from Table 3, in all these cases, the direction of the UID effects are not in the expected direction, *i.e.*, reference sentences (involving non-canonical DO/IO-subject-verb orders) display spikes and troughs in their lexical and syntactic surprisal values.

This result connects directly to prior work (Maurits et al., 2010), which makes a prediction that languages with object-first orders are non-optimal in ensuring an even spread of information across the entire sentence. They define a toy language consisting of only permutations of three words (viz., subject, object and verb). Then they create data for this toy language using English and Japanese child-directed speech obtained from the CHILDES corpus. Subsequently, they demonstrate that in object-first orders, the first word (*i.e.*, the object) is associated with a disproportionate quantum of information because

objects tend to predict ensuing subjects and verbs very accurately. Subsequent words (especially the final verb) are thus rendered to be very uninformative, resulting in a significant trough after the object. For example, the object *water*, restricts predictions related to verbs to a few possibilities like *drink*. In contrast, encountering a verb like *drink* first can trigger multiple object candidates like *water*, *juice* or *tea*. Our own written Hindi data is very different from the toy language created out of the child-directed speech data. Yet, the aforementioned pattern of spikes/troughs prior to the verb is attested in our data as exemplified in the reference-variant pair of sentences below:

(1) a. POTA kanoon-ko pichle raajag sarakaar-ne aatankavaad-se nipatane va aatankee gatividhiyon-par
POTA law-ACC previous central government-ERG terrorism-OBL tackle and terrorist activities-LOC
lagaam-ke liye laagoo kiya tha.
restrain-PSP imposed

The POTA law had been implemented by the previous central government for tackling terrorism and restraining terrorist activities.

b. pichle raajag sarakaar-ne POTA kanoon-ko aatankavaad-se nipatane va aatankee gatividhiyon-par lagaam-ke liye laagoo kiya tha

Here, the reference sentence with object fronting (Example 1a above) has slightly higher lexical surprisal (*i.e.*, higher processing cost) of 41.92 bits compared to the variant (Example 1b with canonical ordering) having lexical surprisal of 41.55 bits. In this case, adding the local UID features (syntactic and lexical) to a model containing lexical surprisal, helps the combined model offset this disadvantage of higher surprisal associated with the reference sentence (in comparison to the variant) and select it. Figure 1 in the Appendix shows the lexical information density changes across the referent-variant pair shown above. In the above examples, in the reference sentence, the first word *POTA* (acronym for Prevention of Terrorism Act) has a higher information density value of 4.5 bits compared to the first word *pichle* (adjective meaning *previous*) in the variant (3.7 bits). However, the acronym is predictive of the word *kanoon* (law), which thus has a low information density value of 1.6 bits, resulting in a trough in the reference sentence. Further research needs to be conducted in order to investigate the information theoretic properties of words belonging to different semantic classes. The above examples also reveal a major lacuna in our current surprisal measures. They do not factor in extra-sentential information going beyond the local lexical and syntactic context. Thus, a word might have a very low probability (higher surprisal) in a particular two-word or local syntactic context, but it might have been mentioned previously in one of the preceding sentences in the discourse context. In Example 1a, the first word (acronym *POTA*) has a high information density value (*i.e.* low trigram probability), but is actually mentioned two sentences before in the preceding context. More generally, out of 13,274 sentences in the entire HUTB, 71.20% sentences contain atleast one content word which is mentioned in the preceding sentence. Persistence effects in language production are a well studied phenomenon (Szmrecsanyi, 2005) and in future we intend to deploy richer models of surprisal estimates incorporating discourse context. One would also expect factors such as the syntactic form of a sentence, its length, focus, or the topic addressed to play a major role in the distribution of information density. These can also be integrated into our models.

5.3 Choice Points in Language Production: Constituent Boundaries

In our UID measures (defined in Section 3) we have made the crucial assumption that individual words are the 'grain size' over which a speaker will spread the information to be transmitted uniformly. While word-based incrementality is taken as standard for language comprehension, language production might exhibit constituent-level incrementality as suggested by psycholinguistic evidence presented by (Hildebrandt et al., 1999). Given that we might often pause at chunk boundaries, this may be effectively allowing for a lowering of information density *in time*. Also, it could be the case that producers are using these spikes to demarcate constituent boundaries.

In order to investigate the above hypothesis, we performed classification experiments using UID measures (both lexical and syntactic) based on constituent boundaries in order to distinguish between corpus and non-corpus sentences. We computed values of constituent-based UID features by plugging in values of information density of the first words of each constituent into the formulae described in Section 2. These new UID features also do not result in significant gains in classification accuracy over and above

lexical surprisal as shown in Table 1 (far right column). The individual performance of constituent-based UID measures are also much worse than the corresponding figures involving the all-words UID measures. The direction of the lexical UID features also suggest the anti-UID effect evinced in the case of the word-based UID measures discussed previously. All these results suggests that UID (as quantified by us) does not shape word order choices in Hindi. We now turn to a discussion of possible theoretical reasons for this.

6 Discussion

In recent years, the UID hypothesis has gained lot of attention as a cognitively plausible account of syntactic reduction phenomena (Levy and Jaeger, 2007; Jaeger, 2010) as well as an explanation for the distribution of various types of word order patterns in language (Maurits et al., 2010). However, our results call into question the role of UID as a predictor of word order choices in Hindi. In this section, we elaborate on various reasons for this.

Ferrer-i-Cancho (2017) establishes that the UID hypothesis is a particular case of the Constant Entropy Rate (CER) hypothesis stated in Genzel and Charniak (2002) and provides a mathematical critique of CER (and hence UID) as applied to word order. The crux of Ferrer-i-Cancho's argument is that for predicting the next element in a sequence, CER and UID are applicable for periodic sequences (the best case in terms of predictability, where a block is repeated as in *abcabcabc...*) as well as sequences of independent identically distributed (*i.i.d.*) elements (the worst case). *i.i.d.* sequences can be random sequences (like scrambled texts) or perfectly homogeneous sequences (example *aaaa...*). Thus, Ferrer-i-Cancho (2017) refutes both CER and derivate UID hypotheses as principles explaining word order on the grounds that these hold for sequences that do not have any kind of order. As a consequence, CER (and UID) cannot be defining characteristics of real texts. Ferrer-i-Cancho (2017) also explains how a modern theory of language and word order in particular consist of a collection of well-established principles and their interactions. Notably, different word order principles are often in conflict with one another. Thus, anti-UID effects are only to be expected. Ferrer-i-Cancho et al. (2013) discuss how probabilities and conditional entropies of natural language might potentially be competing principles with one favouring UID while the other is working against it. Recently, we performed similar experiments on English using syntactic choice data from WSJ and Brown corpora used in Rajkumar et al. (2016). In English also, preliminary results indicate that our UID measures do not significantly improve upon the performance of lexical and syntactic surprisal. This leads further credence to the critique of UID presented above.

Ferrer-i-Cancho (2017) further discusses the empirical success of UID in accounting for syntactic reduction phenomena by showing that reduction is a special case of the principle of compression of codes in standard information theory. Higher order compression allows for codes of length 0, viz. full reduction as in the case of *that*-omission in complement clauses (Jaeger, 2010). First order compressions involve codes of length greater than zero as in the case of contractions like *he's* (instead of the full form *he is*) explained using UID in Frank and Jaeger (2008). Thus, our own empirical results and the recent critique of UID in the literature suggest that while UID might be effective in explaining syntactic reduction phenomena in natural language, its contribution towards a theory of word order is doubtful.

7 Conclusions and Future work[2]

Our results suggest that the UID hypothesis for word order (as quantified by our UID measures) does not shape word order choices in Hindi. Our experiments reveal that these UID measures do not contribute over and above lexical surprisal, a control factor, for predicting the corpus sentence. Moreover, anti-UID effects are attested in the case of object fronting, constructions known to be not favourable to distributing the information uniformly across the utterance. In order to model word order, in the near future we plan to test the efficacy of discourse-context enhanced surprisal estimated using more advanced models like RNNs and LSTMs. We also intend to explore other measures of variation like the *coefficient of variation*, and test our hypotheses on typologically diverse languages from South Asia.

[2]We are grateful to Florian Jaeger and the anonymous reviewers of this workshop and CMCL-2018 for their feedback. The fourth author acknowledges support from IISER Bhopal's Faculty Initiation Grant (IISERB/R&D/2018-19/77).

References

Rama Kant Agnihotri. 2007. *Hindi: An Essential Grammar*. Essential Grammars. Routledge.

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

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

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

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

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681, Oct.

August Fenk and Gertraud Fenk-Oczlon. 1980. Konstanz im kurzzeitgedchtnis - konstanz im sprachlichen informationsflu? *Zeitschrift fr experimentelle und angewandte Psychologie*, 27:400–414, 01.

Gertraud Fenk-Oczlon. 2001. Familiarity, information flow, and linguistic form. In J.L. Bybee and P.J. Hopper, editors, *Frequency and the Emergence of Linguistic Structure*, volume 45, pages 431–448. John Benjamins Publishing Company, 01.

Ramon Ferrer-i-Cancho, ukasz Dbowski, and Fermn Moscoso del Prado Martn. 2013. Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(07):L07001.

Ramon Ferrer-i-Cancho. 2017. The placement of the head that maximizes predictability. an information theoretic approach. *Glottometrics*, 39:38–71, 05.

A. Frank and T.F. Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Cogsci. Washington, DC: CogSci*.

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

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

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

Bernd Hildebrandt, Hans-Jürgen Eikmeyer, Gert Rickheit, and Petra Weiß. 1999. Inkrementelle sprachrezeption. In Ipke Wachsmuth and Bernhard Jung, editors, *KogWis99: Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft*, pages 19–24. Bielefeld University.

T. Florian Jaeger and Esteban Buz. in press. Signal reduction and linguistic encoding. In Eva M. Fernandez and Helen Smith Cairns, editors, *Handbook of Psycholinguistics*, page To appear. Wiley-Blackwell.

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

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

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

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA.

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

Luke Maurits, Dan Navarro, and Amy Perfors. 2010. Why are some word orders more common than others? a uniform information density account. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1585–1593.

K.P. Mohanan and Tara Mohanan. 1994. Issues in word order in south asian languages: Enriched phrase structure or multidimensionality? In Miriam Butt, Tracy Holloway King, and Gillian Ramchand, editors, *Theoretical perspectives on word order in South Asian languages*, pages 153–184. Center for the Study of Language and Information, Stanford, CA.

Günter Neumann and Gertjan van Noord. 1992. Self-monitoring with reversible grammars. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 700–706, Nantes, France. Association for Computational Linguistics.

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

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

C. E. Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27.

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

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

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

A Appendix

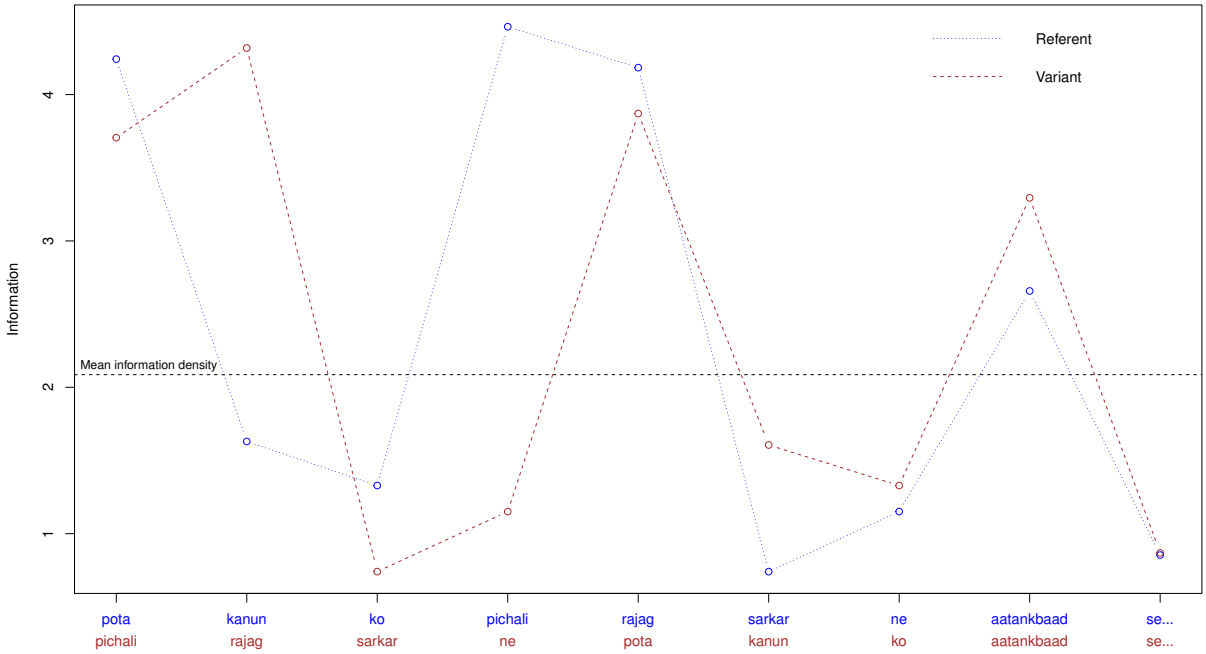

Figure 1: Information variation in bits/word across a pair of reference-variant sentences