Uniform Information Density Effects on Syntactic Choice in Hindi and English

Vishal Singh* IIT Delhi

Ayush Jain* IIT Delhi

Sumeet Agarwal IIT Delhi

Rajakrishnan PR IIT Delhi

1 Introduction

The idea that language, its form and use is optimized to be more efficient dates back to Zipf's 1929 pioneering work which observed that frequent words tend to be shorter in form. More recent work suggests that word length is strongly correlated with the predictability with which it occurs in that context (Piantadosi et al., 2009). Even the use of language has been shown to have such an effect. Instances of the same word which have greater predictability in that context tend to be spoken faster and with less emphasis on acoustic details (Aylett and Turk, 2004; Bell et al., 2003, 2009; Pluymaekers et al., 2005). This raises the possibility that human language production could be organized in terms of processing and communicative efficiency at all levels.

The Uniform Information Density (henceforth UID) hypothesis proposed by Florian Jaeger and colleagues (Frank and Jaeger, 2008; Jaeger, 2010) states that language production exhibits a preference for distributing information uniformly across a linguistic signal. According to this hypothesis, speakers tend to distribute information density across the signal uniformly while producing language, either by omitting optional markers or by explicitly mentioning them. In contrast to the prior work cited above, which looks at information density at *particular choice points* in language production, we examine a variant of the UID hypothesis (as stated above) in the case of *entire sentences* created by syntactic alternations[1].

In this work, we examine the impact of information variance on predicting syntactic choice in Hindi and English. This is the first work on the Hindi language (to the best of our knowledge), which studies the information-theoretic properties pertaining to syntactic choice. Hindi is an SOV language with relatively flexible word order as illustrated by the examples from Mohanan and Mohanan (1994):

(1) a. aaj maa-ne bacce-se *kitaab* *padhne-ko* kahaa
today mother-ERG child-ACC book-NOM read-INF told
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

The meaning expressed by Example (1a) above can be conveyed by any of the sentences in Examples (1b- 1e) obtained by changing the order of constituents of the original sentence[2]. Similarly, our English data also consists of corpus sentences and their syntactic choice variants.

More generally, testing such hypotheses on a typologically diverse set of languages is imperative to advance theories of language production as results from previously studied languages need not reflect universal processing mechanisms. As (Jaeger and Norcliffe, 2009) compellingly argue, cross-linguistic inquiries of language production are still rare, but are imperative for refining existing theories in addition to validating or disproving current hypotheses.

Here, we estimate the information density of sentences using both lexical and part-of-speech (POS) trigram models. Based on these estimates, we model the uniformity of information across a sentence by proposing five distinct UID measures. Our experiments primarily involved the task of classifying Hindi and English data into reference sentences (like Example 1a above) and artificial

[1]Please refer to (i Cancho et al., 2013) inter-alia, for a detailed discussion of the idea that choosing the better variance of information density at specific choice points is not the same as having overall lower than expected variance in information density.

[2]These variants involve marked intonational patterns to facilitate comprehension as noted by Mohanan and Mohanan (1994)

variants (say Examples 1b- 1e) which were created by linearizing dependency graphs corresponding to reference sentences (obtained from standard corpora) in order to create syntactic choice variants thereof. The UID measures alluded to above were deployed as features in machine learning models to perform the task of binary classification. Our results indicate that for Hindi, our UID measures do not help our SVM model in predicting the corpus choice sentence over above the word and POS-based trigram models. In the case of English, the UID measures have a slight impact in improving SVM classification accuracy over the trigram model baseline. Thus we conclude that our measures, based on our version of the UID hypothesis to model the uniformity of information spread in the language signal, is not a robust predictor of syntactic choice in Hindi and in English, POS-based UID exhibits a weak effect.

2 UID Measures

The UNIFORM INFORMATION DENSITY principle discussed by Jaeger (2010) predicts that language production is optimized to distribute information uniformly across the utterance. Here we define the UID measures 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 Collins (2014) and their normalized counterparts are our own. N: number of words in a sentence; w_i: i^{th} word of a sentence; ng_i: n-gram information density (negative log-prob) of the i^{th} word of the sentence; μ: mean n-gram information density of the sentence, i.e., $\mu \equiv \frac{1}{N}\sum_{i=1}^{N} ng_i$.

$$UIDglob = -\frac{1}{N}\sum_{i=1}^{N}(ng_i - \mu)^2$$
$$UIDloc = -\frac{1}{N}\sum_{i=2}^{N}(ng_i - ng_{i-1})^2$$
$$UIDglobNorm = -\frac{1}{N}\sum_{i=1}^{N}(\frac{ng_i}{\mu} - 1)^2$$
$$UIDlocNorm = -\frac{1}{N}\frac{\sum_{i=2}^{N}(ng_i - ng_{i-1})^2}{\mu^2}$$
$$UIDlocLocNorm = -\frac{1}{N}\sum_{i=2}^{N}(\frac{ng_i}{ng_{i-1}} - 1)^2$$

3 Experiments and Discussion

3.1 Ranking Experiments

Trigram information density and all UID measures were computed for all corpus sentences and their variants. The corpora used are the Hindi-Urdu Treebank (Bhatt et al., 2009) and the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993). The corpus sentences and variants were ranked using these measures to see if they tend to pick out the corpus choice. We would expect the corpus sentence to be ranked near the top, if the UID hypothesis (as quantified via our measures) holds. Figure 1 shows the results on Hindi. Our normalized UID measures (which should be uncorrelated with the overall trigram information density) show no tendency to pick out the actually-occurring word order; surprisingly, they show the opposite tendency at times! Our English results (not shown) confirm similar findings obtained by other researchers[3] that there is no evidence that the variance of Shannon information across words within sentences is lower than expected by chance.

The analysis was repeated with the trigram models being trained on the POS tag sequences and all features being computed using these models (Figures 2 and 3). This shows a substantive difference between Hindi and English. English, to some degree, appears to be following our version of the UID hypothesis when generalizations over words (POS tags) are factored in. This is in accordance with previous work on English syntactic alternations (Collins, 2014), where the UID measures proposed in that work were significant predictors of human ratings of sentence quality. However, for Hindi we find no indication of UID governing production choice. Also, the more global measures seem to better predict production choice for English; for the more local ones, we still see the mysterious spike at the bottom rank, i.e., a bunch of corpus sentences with higher local information variation than any of their variants!

3.2 Pairwise Classification using SVMs

The rank analysis only considers one feature at a time. To examine the extent to which different features might complement each other in determining production choice, we used linear Support Vector Machines (SVMs) for the binary classification task of corpus sentences vs. non-corpus variants.

Since the data sets have many more non-corpus than corpus variants, we use a technique from Joachims (2002) to convert it into a balanced setting (see Rajkumar et al., 2016 for details). The binary classification task is then to identify each given pair's type, i.e., given such a pair, identify

[3] Florian Jaeger (p.c.) based on results from American English (written and spoken), German, Arabic (Modern Standard), Czech and Mandarin Chinese documented in Gildea and Jaeger (Submitted)

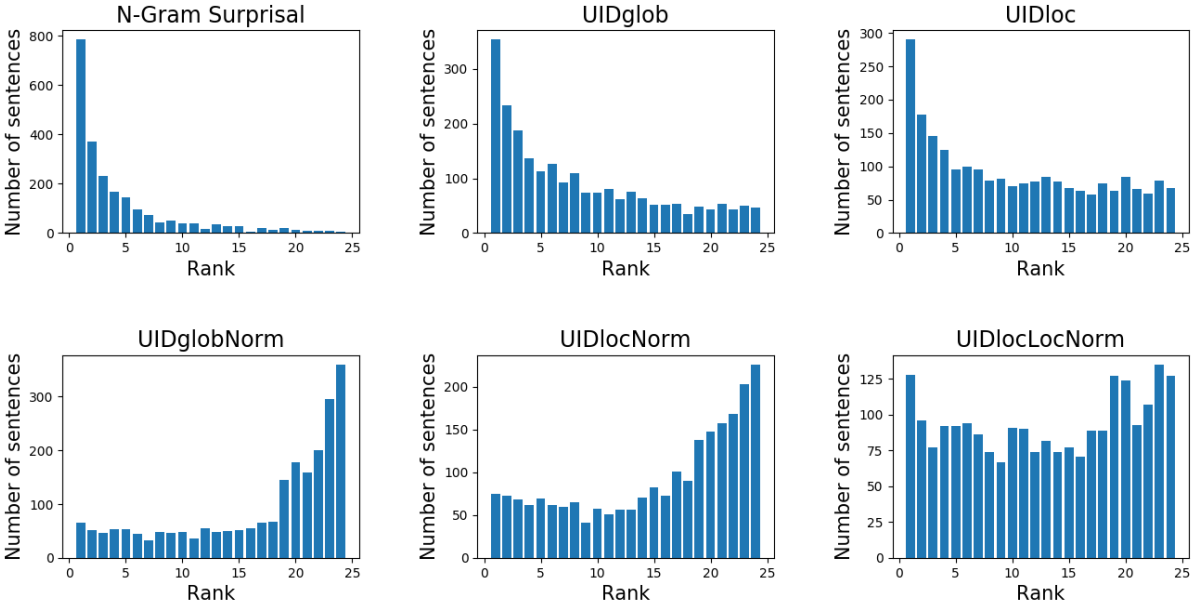

Figure 1: Rank histograms, depicting the rank of the actual corpus sentence amongst all its variants, based on measures computed using lexical trigram models (Hindi: 2256 reference sentences each having 23 variants)

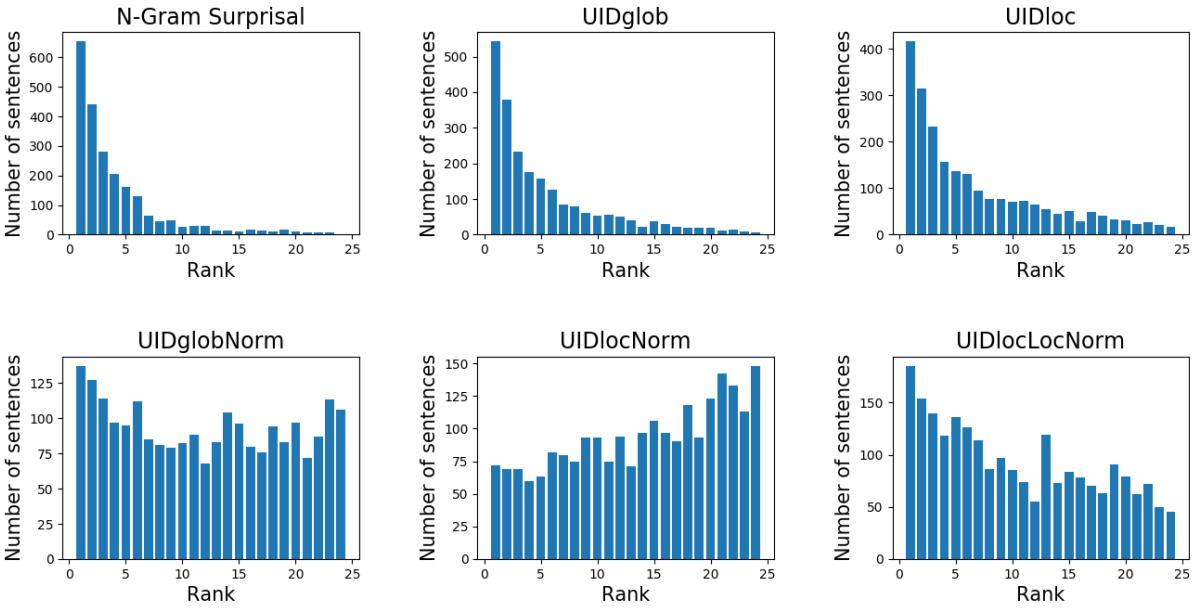

Figure 2: Rank histograms, depicting the rank of the actual corpus sentence amongst all its variants, based on measures computed using POS trigram models (Hindi: 2256 reference sentences each having 23 variants)

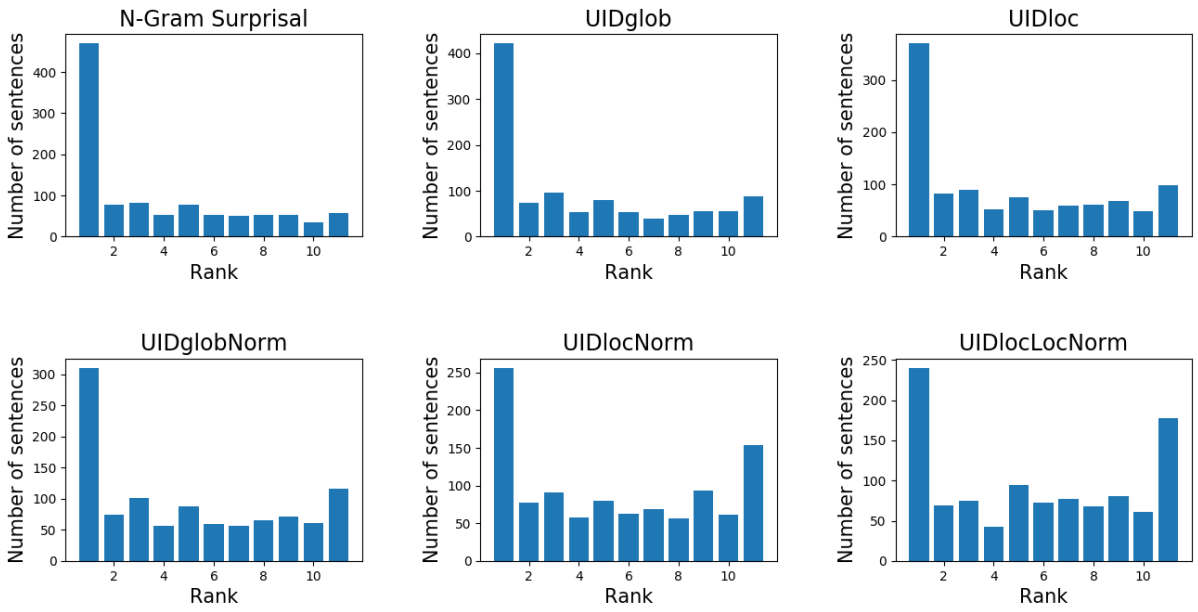

Figure 3: Rank histograms, depicting the rank of the actual corpus sentence amongst all its variants, based on measures computed using POS trigram models (English: 1060 reference sentences each having 10 variants)

Features	**Accuracy**	
	Hindi (51888 datapoints)	**English** (10600 datapoints)
$3g.ID$	**89.78**%	78.08%
$3g.ID + UIDglob + UIDloc$	84.08%	**78.37**%
$3g.ID + UIDglobNorm + UIDlocNorm + UIDlocLocNorm$	89.10%	77.96%
$3g.ID + All\ UID$	79.11%	78.34%

Table 1: SVM Ranking results. For all features indicated, two versions are included, one based on lexical trigram models and the other based on part-of-speech tag trigram models. $3g.ID$ is the overall trigram information density.

whether the corpus sentence is the first one or the second one. Table 3.2 shows the classification results for models trained on different subsets of our features.

The addition of normalized UID measures for both languages resulted in lower accuracy than the baseline (using only trigram information density features). This indicates that these measures are not adding anything useful beyond overall trigram information density. However, when adding only the unnormalized UID measures, we see a slight increase from the baseline for English of about 0.3%. Hindi however shows a substantive drop in this case, suggesting that our UID measures are not predictive of production choice for Hindi and are in fact confusing the classifier. Overall, UID as quantified by us appears to have at best a slight effect on determining word order for English, and none at all for Hindi.

4 Conclusions

The version of the UID hypothesis for word order described in this paper and subsequently quantified using our UID measures does not seem to shape word order choices in Hindi, unlike other processing principles. For English, after controlling for another strong factor like n-gram information density, UID has a very weak effect on syntactic choice, and the main observation is a clear difference between lexical and POS-based versions of our UID measures. This seems to agree with our UID hypothesis as operating at a level of granularity above words for English. However, for Hindi, even the POS-based versions don't seem to be informative at all about word-order choice. Does this imply that UID is somehow a much weaker constraint for Hindi production, if at all? Hindi is a verb-final language with extensive case-marking and flexible word-order: so it may also be that the notions of information and information density need to be defined differently for such a language, compared to English.

References

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1):31–56.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language* 60(1):92–111.

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America* 113(2):1001–1024.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL-IJCNLP '09, pages 186–189. http://dl.acm.org/citation.cfm?id=1698381.1698417.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research* 43(5):651–681. https://doi.org/10.1007/s10936-013-9273-3.

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

Daniel Gildea and Florian Jaeger. Submitted. Language structure shaped by the brain: Human languages order information density .

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. http://stacks.iop.org/1742-5468/2013/i=07/a=L07001.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage information density. *Cognitive Psychology* 61(1):23–62. http://dx.doi.org/10.1016/j.cogpsych.2010.02.002.

T. Florian Jaeger and Elizabeth Norcliffe. 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. http://dx.doi.org/10.1111/j.1749-818X.2009.00147.x.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '02, pages 133–142. https://doi.org/10.1145/775047.775067.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.* 19(2):313–330. http://dl.acm.org/citation.cfm?id=972470.972475.

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*, Center for the Study of Language and Information, Stanford, CA, pages 153–184.

Steven T Piantadosi, Harry J Tily, and Edward Gibson. 2009. The communicative lexicon hypothesis. In *The 31st annual meeting of the Cognitive Science Society*.

Mark Pluymaekers, Mirjam Ernestus, and R Harald Baayen. 2005. Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America* 118(4):2561–2569.

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.

George Kingsley Zipf. 1929. Relative frequency as a determinant of phonetic change. *Harvard studies in classical philology* 40:1–95.