Effects of Uniform Information Density in English Syntactic Choice

Sidharth Ranjan[1] Rajakrishnan Rajkumar[2] Sumeet Agarwal[1]
IIT Delhi[1], IISER Bhopal[2]
sidharth.ranjan03@gmail.com, rajak@iiserb.ac.in, sumeet@iitd.ac.in

According to the Uniform Information Density (henceforth, UID) hypothesis (Levy and Jaeger, 2007; Jaeger, 2010), language production shows a preference to distribute information uniformly across the linguistic signal. The aforementioned studies validated the UID hypothesis using syntactic reduction phenomena, *viz.*, *that*-complementizer and relativizer mention (or omission), at *particular choice points* in a sentence. In contrast, we studied the effects of UID across *entire sentences* in English syntactic choice alternations. We used a multi-construction dataset (dative alternation, quotation, pre and postverbal adjuncts) created by Rajkumar et al. (2016, see Section 4 and Table 3 therein), consisting of 6415 Brown corpus reference sentences and 8385 automatically constructed variants expressing the same idea using a different order of words:

(1) *She came to New York from Detroit as a teenager but with a sponsor instead of a chaperone.* vs *She came* **from Detroit** *to New York as a teenager but ...*

Collins (2014) showed that measures modelling UID were significant predictors of human acceptability ratings of syntactic choice. Inspired from this work, we quantified uniformity in information spread in a sentence using the following measures defined at both lexical and syntactic levels: *coefficient of variation (COV), global and local UID measures (UIDglob and UIDloc), normalized global and local UID measures (UIDglobNorm and UIDlocNorm)* (see final page for formulae). For each reference and variant sentence, we computed various UID measures using per-word estimates of trigram and PCFG surprisal (see final page for surprisal estimation procedure). Sentence-level surprisal values were obtained by summing the per-word surprisal scores over all the words in a sentence. Since our original dataset is highly unbalanced (variants outnumber reference sentences), we used a transformation (see final page) to pair reference sentences with each of its variants (Joachims, 2002). Subsequently surprisal and UID measures were incorporated into a logistic regression classifier aimed to distinguish reference sentences from variants. The output of the regression model containing individual predictors showed a consistent negative regression coefficient for all our predictors including surprisal and UID measures except syntactic UIDglob (see Table 1 overleaf). The negative regression weights of surprisal predictors suggest that the log odds of predicting a reference sentence (over a variant) increases with lower surprisal values. The negative coefficient for our UID measures indicates that reference sentences (preferred choice) have lower uniformity as compared to variants suggesting a violation of UID predictions. Further, none of the UID measures discernably improved reference sentence prediction accuracy over and above a baseline model containing trigram and PCFG surprisal (Table 2). This seems to suggest a null effect of UID in presence of surprisal-based controls, contrary to previous work on UID and word order cited above.

Construction-specific analyses revealed a significant UID effect of both UIDglobNorm and COV lexical measures for the dative alternation construction. These measures significantly contributed over and above trigram and PCFG surprisal in predicting reference sentences with postverbal NP-NP structure over variants with NP-PP structure. Further exploration showed that the success of UID in dative alternation was mostly associated with NP-NP structures involving pronouns (76% cases). Figure 1 depicts the trend visually for the examples below:

(2) *She wants to pay [you]$_{NP}$ [a visit]$_{NP}$.* vs *She wants to pay [a visit]$_{NP}$ [to you]$_{PP}$.*

This result suggests that pronouns might carry peculiar information properties, which requires further investigation. Overall, we conclude that UID might be effective for accounting syntactic reduction phenomena but not for predicting reference sentences amidst variants. As a part of future work, we plan to create counterfactual treebanks along the lines of Hahn et al. (2020) and test UID predictions for syntactic choice phenomena by training language models on them.

Predictor(s)	Logistic Regression			
	Lexical UID		*Syntactic UID*	
	Weight(s)	%Acc	Weight(s)	%Acc
3g surprisal	-0.66	65.89	-0.66	65.89
PCFG surprisal	-0.43	72.27	-0.43	72.27
UIDglob	-0.86	53.89	0.11	52.63
UIDloc	-0.1	52.15	-0.002	49.58
UIDglobNorm	-12.43	59.58	-0.72	51.13
UIDlocNorm	-2.41	56.92	-0.69	52.37
COV	-13.60	59.44	-0.77	51.03

Table 1: Performance of logistic regression models consisting 8385 data points

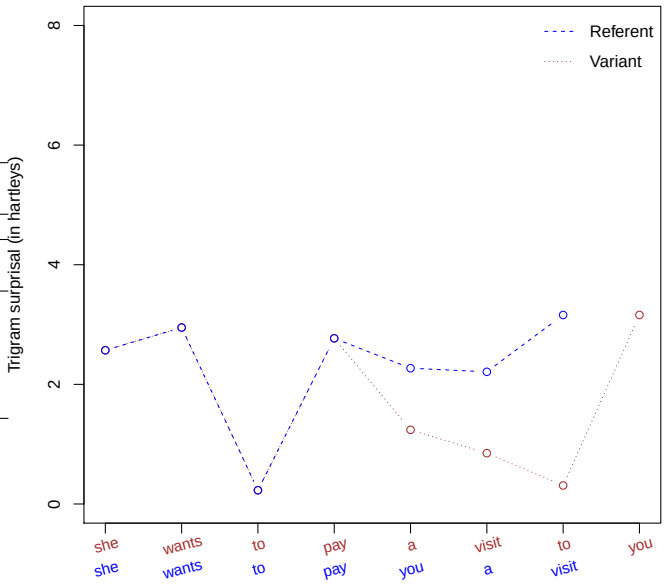

Figure 1: Trigram information variation (hartleys/word) across a pair of reference-variant sentences

Predictor(s)	Logistic Regression			
	Lexical UID measures		*Syntactic UID measures*	
	Weight(s)	%Accuracy	Weight(s)	%Accuracy
Baseline=Trigram+PCFG surprisal	-0.38 -0.37	73.99	-0.38 -0.37	73.99
Baseline+UIDglob	-0.39 -0.37 0.19	74.00	-0.38 -0.37 0.03	74.00
Baseline+UIDloc	-0.38 -0.37 -0.08	73.93	-0.38 -0.37 -0.04	73.91
Baseline+UIDglobNorm	-0.42 -0.36 2.09	73.92	-0.38 -0.37 -0.11	73.94
Baseline+UIDlocNorm	-0.37 -0.37 -0.25	73.89	-0.38 -0.37 -0.41	73.89
Baseline+COV	-0.42 -0.36 2.67	73.92	-0.38 -0.37 0.05	73.97

Table 2: Performance of logistic regression models consisting of 8385 data points

References

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

Hahn, M., Jurafsky, D., and Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353.

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

Jain, A., Singh, V., Ranjan, S., Rajkumar, R., and Agarwal, S. (2018). Uniform Information Density Effects on Syntactic Choice in Hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48, Santa Fe, New-Mexico.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD*, KDD '02, pages 133–142, New York, USA. ACM.

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

Levy, R. P. (2018). Communicative efficiency, uniform information density, and the rational speech act theory.

Rajkumar, R., van Schijndel, M., White, M., and Schuler, W. (2016). Investigating locality effects and surprisal in written english syntactic choice phenomena. *Cognition*, 155:204–232.

Supplementary Materials

We have used sentence-level UID measures as proposed in Jain et al. (2018) for investigating the effects of UID on Hindi syntactic word-ordering choice. The information density of a sentence is computed via per-word contextual probabilities using trigram and PCFG parser surprisal models as explained below:

Notation: N is the number of words in a sentence, id_i is the information density or surprisal (negative lexical/syntactic log-probability) of the i^{th} word of the sentence, μ is defined as the mean information density of the sentence, *i.e.*, $\mu \equiv \frac{1}{N}\sum_{i=1}^{N} id_i$ and σ is spread of information from the mean, *i.e.*, $\sigma \equiv \sqrt{\frac{\sum_{i=1}^{N}(id_i-\mu)^2}{N}}$

1. **Coefficient of Variation**: $COV = -\frac{\sigma}{\mu}$
   This captures the degree of variation (extent of variability) among different data series. The negative sign indicates higher uniformity corresponding to the lower ratio of the standard deviation to mean.
2. **Global UID Measure**: $UIDglob = -\frac{1}{N}\sum_{i=1}^{N}(id_i-\mu)^2$
   This captures the uniformity in information (negative variance) across an entire sentence (globally). The negative sign corresponds to minimization of variance of information, *viz.*, different points in a sentence should carry similar information.
3. **Local UID Measure**: $UIDloc = -\frac{1}{N}\sum_{i=2}^{N}(id_i-id_{i-1})^2$
   This measure captures the uniformity in information (negative mean-squared variance) per word relative to the previous word (locally).
4. **Normalized Global UID Measure**: $UIDglobNorm = -\frac{1}{N}\sum_{i=1}^{N}(\frac{id_i}{\mu}-1)^2$
   This measure encapsulates the global variance of information as explained in $UIDglob$ above, but is also normalized relative to the mean information density (μ) over all words in the sentence.
5. **Normalized Local UID Measure**: $UIDlocNorm = -\frac{1}{N}\frac{\sum_{i=2}^{N}(id_i-id_{i-1})^2}{\mu^2}$
   This measure also normalises $UIDloc$ viz., local information using the mean information density of the sentence.

We estimate trigram surprisal by training language model on Open American National Corpus (OANC) containing 1 million sentences of mixed genres. We compute PCFG surprisal by training Berkeley constituency parser on WSJ sections (02-21) of the Penn Treebank (PTB) corpus. In future inquiries, we plan to test UID predictions by incorporating string complexity over syllables and graphemes in UID metric computation as proposed in Levy (2018, see Equation 6 therein).

$$\textit{choice} \sim \textit{ngram surprisal} + \textit{parser surprisal} + \textit{UID measures} \quad (1)$$

All the experiments have been conducted in R using Glm package in 10-fold cross-validation setting. In Equation 1, the variable *choice* is a binary dependent variable (1 stands for correct choice i.e., corpus sentence is ranked higher than its paired variant and 0 denotes the incorrect choice i.e., model prefers variant over its reference sentence). We transformed our data set using a technique originally proposed by Joachims (2002) for ranking web pages. 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. We trained a machine learning model on the difference between the aforementioned feature vectors as per the equations below:

$$1.\ \mathbf{w}\cdot\phi(Reference) > \mathbf{w}\cdot\phi(Variant) \qquad 2.\ \mathbf{w}\cdot(\phi(Reference)-\phi(Variant)) > 0$$

Equation 1 above shows the model prediction where reference sentence outranks one of its variants when the dot product of the feature vector of the reference sentence and $\mathbf{w}$ (learned feature weights) is greater than the corresponding dot product of the variant sentence. This relationship can also be expressed in the form of Equation 2 , where the feature values of the first member of the pair were subtracted from the corresponding values of the second member. We created ordered pairs consisting of the feature vectors of reference-variant sentences.