Title
Information density of encodings: The role of syntactic variation in comprehension

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

Authors
Sikos, Les
Greenberg, Clayton
Drenhaus, Heiner
et al.

Publication Date
2017

Peer reviewed

Information density of encodings: The role of syntactic variation in comprehension

Les Sikos (sikos@coli.uni-saarland.de), Clayton Greenberg (claytong@coli.uni-saarland.de), Heiner Drenhaus (drenhaus@coli.uni-saarland.de) and Matthew W. Crocker (crocker@coli.uni-sb.de)
Department of Language Science and Technology, Saarland University, 66123 Saarbrücken, Germany

Abstract

The Uniform Information Density (UID) hypothesis links production strategies with comprehension processes, predicting that speakers will utilize flexibility in encoding in order to increase uniformity in the rate of information transmission, as measured by surprisal (Jaeger, 2010). Evidence in support of UID comes primarily from studies focusing on word-level effects, e.g. demonstrating that surprisal predicts the omission/inclusion of optional words. Here we investigate whether comprehenders are sensitive to the information density of alternative encodings that are more syntactically complex. We manipulated the syntactic encoding of complex noun phrases in German via meaning-preserving pre-nominal and post-nominal modification in contexts that were either predictive or non-predictive. We then used the G-maze reading task to measure online comprehension during self-paced reading. Results were consistent with the UID hypothesis. In predictive contexts, post-nominal encodings elicited a more uniform distribution of processing effort. Conversely, in non-predictive contexts, more uniform effort was found for pre-nominal encodings.

Keywords: Language comprehension; surprisal; uniform information density hypothesis; G-maze; self-paced reading.

Introduction

Levy and Jaeger's (2007) Uniform Information Density hypothesis postulates that speakers adjust their lexical and syntactic realization of a message for the benefit of comprehenders. Specifically, they suggest that there is an overarching preference to produce message encodings that distribute information as evenly as possible across the linguistic signal. This account fundamentally links encoding and decoding processes, asserting that language producers will exploit the flexibility in encoding so as to increase uniformity in the rate of information transmission, as measured by surprisal (Hale, 2001; Levy, 2008). As such, the UID hypothesis can be viewed as part of a rational theory of communication — from an information theoretic perspective — in which encoding strategies take into account resource limitations of the comprehender.

There is robust empirical evidence that surprisal accounts for cognitive load during comprehension — at least at the level of individual words in a sentence (Drieghe, Rayner & Pollatsek, 2005; Kliegl, Grabner, Rolfs & Engbert, 2004; Rayner, Aschby, Pollatsek & Reichle, 2004; Rayner & Well, 1996). However, there exists little direct online evidence that comprehenders are indeed sensitive to the surprisal and density profiles of *alternative syntactic encodings* — a critical assumption underlying the UID hypothesis. Furthermore, current support for UID in *production* is limited to relatively local encoding choices, such as *that*-deletion (Jaeger, 2010; Levy & Jaeger, 2007), contraction (Frank & Jaeger, 2008), and the use of single word equivalents that vary in word length (*chimpanzee* vs. *chimp*; Mahowald, Fedorenko, Piantadosi & Gibson, 2013). Although the above studies provide important and compelling support for the notion that UID modulates aspects of syntactic encoding, the generality of the findings is limited by the observation that all the phenomena are instances of highly local syntactic reduction.

The goal of the current study is to investigate whether comprehenders are sensitive to the information density of more complex alternative syntactic encodings. Consider the following examples:

(1) The journalist published… Predictive context
(2) The man evaluated… Non-predictive context

a. …[the carefully written essay]. Obj NP_{adj}
b. …[the essay that was carefully written]. Obj NP_{rel}

Each object noun phrase (NP) above arguably expresses the same message,[1] however (a) uses a pre-nominal adjective phrase while (b) uses a post-nominal relative clause. While the head noun (*essay*) is more expected in the predictive (1) than non-predictive (2) contexts, the expectation for the adjective *carefully* presumably does not differ across contexts. One potential encoding strategy for increasing the uniformity of information density would be to produce low-surprisal words early in the sentence, as this may facilitate the processing of subsequent less predictable words. For instance, in the non-predictive context, the UID hypothesis predicts a processing advantage for the pre-nominal encoding because *carefully written* should reduce the surprisal of *essay.* In the predictive context, on the other hand, the pre-nominal encoding may result in a trough in information density at *essay,* as it is highly expected (and thus not very informative) following the verb and modifiers. In this case, the post-nominal relative clause may distribute the informational content more uniformly. The UID hypothesis therefore, predicts a greater benefit for the relative clause encoding in more predictive contexts.

We tested the above predictions using a self-paced reading design to measure online differences in cognitive load during the critical object NP.

[1] Choices in linguistic encoding are known to be influenced by aspects of information structure, including contrastive focus,

Experiment

Our primary goal was to test whether comprehenders are sensitive during online processing to differences in the information density of alternative syntactically-complex encodings that nevertheless convey a similar message. The materials crossed two factors (context × encoding), as illustrated in Table 1. Because this manipulation distributes information within the critical region differently across conditions, we assessed cognitive load using a variation of self-paced reading that is less susceptible to spill-over effects than standard forms of self-paced reading. The grammaticality maze task (G-maze; Forster, Guerrera & Elliot, 2009) can precisely identify the word at which processing time differences emerge during online comprehension (Witzel, Witzel & Forster, 2010) and is therefore well-suited for our purposes. In this task sentences are presented word by word as a sequence of forced choices between two alternatives, only one of which continues the sentence grammatically. If the participant successfully navigates the "maze" by choosing the correct word from each pair, the selected words form a coherent sentence (Figure 1).

Methods

Participants Twenty-seven native German speakers (age *M* = 24, *SD* = 2.6) with normal or corrected to normal vision were recruited from the Saarland University community and were compensated 8€ for their participation. Participants that did not successfully navigate at least 70% of mazes in all experimental conditions were excluded (n = 3).

Materials Forty-eight sets of sentences were constructed in German by crossing context (predictive, non-predictive) and syntactic encoding of the object NP (pre-nominal modification, post-nominal modification), resulting in four conditions per item (Table 1). In order to create the context manipulation, the same object noun (e.g., *Essay*, "essay") was used in all conditions, but the object was designed to be more expected in predictive contexts than non-predictive contexts. This was accomplished by choosing different subject–verb combinations for each context. Subject–verb combinations were neutral with respect to the object noun in the non-predictive context (e.g., *Der Mann bewertete*, "The man evaluated"), but were semantically associated with the object in the predictive contexts (e.g., *Der Journalist veröffentlichte*, "The journalist published"). Importantly, however, highly expected object nouns (e.g., *Artikel*, "article") were avoided in order to increase the possibility of detecting surprisal differences between predictive/pre-nominal and predictive/post-nominal conditions.

The information density of object NPs was manipulated via pre- and post-nominal modification, affecting both the linear ordering and length (in words) of the message. Pre-nominal modifiers (e.g., *sorgfältig verfassten*, "carefully written") were shorter, containing 2 to 4 words, but positioned the head noun at the end of the NP. Post-nominal modifiers used a relative clause construction (e.g., *der sorgfältig verfasst worden war*, "that was carefully written") and were therefore longer, ranging from 4 to 6 words, and constrained the head noun to the beginning of the NP. To avoid having any words within the critical object NP region be sentence-final, all items ended with an adverbial phrase.

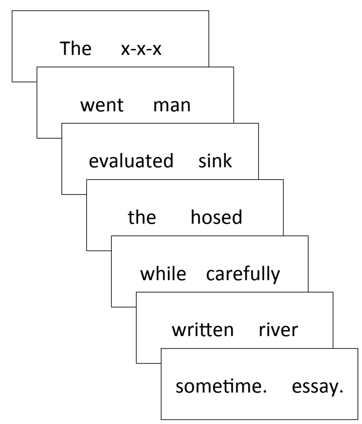

Figure 1: Example trial structure of G-maze task. Sentences (in German) were presented word by word as a sequence of forced choices between two alternatives, only one of which continued the sentence grammatically.

Table 1: Example stimulus item in four conditions with approximate English translations. The critical region of interest was the object NP. RTs were analyzed separately for the object noun (bold) and modification region (underlined).

Context	Encoding	Example
Predictive	Post-nominal	*Der Journalist veröffentlichte den* ***Essay,*** *der sorgfältig verfasst worden war, unter Einbeziehung des größeren Kontextes.* "The journalist published the **essay** that was carefully written, taking into account the larger context."
Predictive	Pre-nominal	*Der Journalist veröffentlichte den sorgfältig verfassten* ***Essay*** *unter Einbeziehung des größeren Kontextes.* "The journalist published the carefully written **essay**, taking into account the larger context."
Non-predictive	Post-nominal	*Der Mann bewertete den* ***Essay,*** *der sorgfältig verfasst worden war, unter Einbeziehung des größeren Kontextes.* "The man evaluated the **essay** that was carefully written, taking into account the larger context."
Non-predictive	Pre-nominal	*Der Mann bewertete den sorgfältig verfassten* ***Essay*** *unter Einbeziehung des größeren Kontextes.* "The man evaluated the carefully written **essay**, taking into account the larger context."

Four counterbalanced lists were constructed from these materials according to a Latin Square design such that each list contained 12 items in each condition, but no item appeared more than once in the same list. An additional 48 sentences with the same structures as above, but containing highly predictable object nouns, were constructed as fillers (e.g., *Der Schneider zerschnitt den stark gemusterten Stoff am Mittwoch.*, "The tailor cut the heavily patterned fabric on Wednesday."). Half of the filler sentences contained pre-nominal modification of the object noun and the other half contained post-nominal modification. No object nouns were repeated across experimental or filler items.

Cloze probability and contextual constraint An offline Cloze completion study was conducted to confirm that object nouns were more expected following predictive than non-predictive contexts, but were not highly expected in either context. A separate group of 58 native German speakers (age M = 22.0, SD = 2.9) were presented with sentence fragments from the 48 experimental items described above. Fragments contained only the contexts, followed by a blank (e.g., *Der Mann bewertete___; Der Journalist veröffentlichte___*). Predictive and non-predictive contexts were counter-balanced across two lists. Participants were asked to fill in the blank with the first determiner–noun combination that came to mind. Cloze probabilities were computed as the percentage of participants who provided the experimental object noun for a particular item. As expected, object nouns had low cloze probabilities in both contexts but were reliably more expected following predictive (cloze = 0.06, SD = 0.18) than non-predictive contexts (cloze = 0.00, SD = 0.01), $t(47) = -2.32$, $p < .05$.

The percentage of the most frequently occurring response to each sentence fragment in the cloze test was also used to assess the contextual constraint of predictive and non-predictive contexts. As expected, the mean constraint of predictive contexts was reliably greater (51.3%) than that of the non-predictive contexts (21.3%), $t(47) = -8.46$, $p < .001$.

Surprisal profiles To compare our response time results against a more theoretical notion of predictability, we computed surprisals for all experimental stimuli using an interpolated modified Kneser-Ney 5-gram language model trained on a 2017-01-01 dump of German Wikipedia. To obtain the corpus, we filtered the original XML dump using the tool WikiExtractor, split the corpus into sentences using the NLTK sentence splitter for German, and preprocessed the resulting dataset.[2] After replacing all types occurring fewer than 15 times with <*unk*>, the vocabulary size was 833,734.[3] We split the corpus into training, development, and test sections with the ratio 8:1:1. The resulting training section contained 666,561,150 tokens. The model was trained using the SRILM toolkit (Stolcke, 2002) and achieved perplexities of 25 on the training section, 201 on the test section, and 1583 on our stimuli.[4]

Procedure Participants were randomly assigned to a stimulus list (6 per list). The G-maze task was implemented in E-prime (Schneider, Eschman, & Zuccolotto, 2002). Each trial began with two crosses (+) that remained on screen for 1000 ms, indicating where subsequent word pairs would appear. Each word in the sentence (except the first word) was then presented together with a foil word,[5] which was not a grammatical continuation of the sentence. The first word in every sentence was paired with "x-x-x". The presentation side (left, right) was randomized such that the correct word appeared equally often on each side. Any punctuation (i.e., comma, period) that appeared with a word also appeared with its foil. Participants were instructed to choose as quickly and as accurately as possible the word that best continued the sentence. Participants indicated their selection by pressing the left or right button on a button box and the amount of time required for selecting the grammatical continuation was recorded as the response time (RT). If the correct word was chosen, the next pair of words appeared automatically. However, if a foil word was selected, negative feedback (*Inkorrekt*, "Incorrect") was displayed and the trial was aborted. Once the end of a sentence was reached, positive feedback (*Korrect*, "Correct") was given. Participants initiated each new trial by button press.

To confirm that participants read the sentences for meaning, a Yes/No comprehension question appeared after 1/3 of the items. Half of the questions asked about the subject noun and half about the object noun. The correct answer was Yes for 50% of questions. Participants used the button box to respond. No feedback was given.

In order to familiarize participants with the task, five practice items (three with comprehension questions) were presented before the experiment began. Participants took approximately 40 minutes to complete the experiment.

[2] Lowercased, replaced punctuation with space, replaced digits with NUM, removed empty lines, replaced tabs with spaces, removed multiple spaces, removed multiple NUMs, replaced umlauts by their conventional character bigrams, and added sentence begin and end markers.

[3] A threshold of 15 was selected since this was the highest possible while maintaining a less than 1% out of vocabulary rate on a different corpus (EUROPARL).

[4] The sharp difference in perplexity scores between the test corpus and stimuli suggests that the German Wikipedia corpus is not an ideal match for our stimuli. We return to this point in the discussion.

[5] Foils were created in a two-stage process. First, a custom Python script randomly selected a foil candidate for each word in each experimental and filler item. Foil candidates were constrained such that they did not appear in bigrams with the correct word at the previous position in the sentence within a large German corpus. Second, each foil was then hand checked by at least two trained native-German linguists to ensure that it was not a grammatical continuation of the sentence. The same foil was used for identical words (or derivationally related words) across conditions.

Results and Discussion

Completed mazes Overall performance on the G-maze task was high, with participants successfully navigating 85.6% (SD = 0.08) of experimental and filler items to completion. However, because the critical region of interest was the object NP, the RT analyses reported below were conducted on all experimental items that were completed through at least the end of the critical region (M = 0.90, SD = 0.06).

Comprehension question accuracy Performance on the comprehension questions was near ceiling (M = 0.97, SD = 0.04), confirming that participants were reading the sentences for meaning during the G-maze task.

Response time RTs were analyzed with linear mixed effects models with participants and items as crossed, independent, random effects. All models included maximal random effects structures (Barr, Levy, Scheepers & Tily, 2013). Analyses were conducted using the *lmer* function (lme4 library, version 1.1-10; Bates & Sarkar, 2007) in the statistics software package R, version 3.2.2 (R Development Core Team, 2013). Fixed effects were evaluated via likelihood ratio tests implemented in *lmerTest* (Kuznetsova, Brockhoff & Christensen, 2015), where denominator *df* was estimated using the Satterthwaite method. We report estimates, standard errors, t and p values associated with likelihood ratio tests for significant results only.

All raw RTs that were abnormally low (below 200 ms) or abnormally high (above 5000 ms) were excluded (0.3%), and outliers exceeding 3 standard deviations by participant were then trimmed (1.8%). The remaining RTs were adjusted for word length (Ferreira & Clifton, 1986) and punctuation using a linear mixed effects regression model with fixed effects for word length, punctuation (i.e., whether or not a comma or period was presented with the word), and their interaction. The residuals of this model, length-adjusted RTs, served as the dependent variable in the analyses reported below.[6] Because the number of words used to modify object nouns varied across items and conditions, we computed the length-adjusted RT for the modification region by averaging across modifier words.

The upper panel of Figure 2 presents the mean length-adjusted word-by-word RTs for each condition. Differences first emerge at the subject noun, where RTs were slower for predictive than non-predictive contexts. This is not surprising as these words (e.g., *journalist*) are less frequent than their non-predictive counterparts (e.g., *man*). More relevant to the research question, all four conditions diverge within the critical object NP region (Table 2).

Object noun analysis. Length-adjusted RTs for object nouns were regressed onto a model including fixed-effect factors for context (predictive, non-predictive), encoding (pre-nominal, post-nominal), and their interaction. In order to control for task adaptation, a main effect of stimulus order was also included.

[6] Qualitatively identical results are obtained when raw RTs are used.

Figure 3 (left panel) shows that object nouns were read more quickly in predictive than non-predictive conditions, β = -161.01, SE = 29.59, $t(44.22) = -5.44$, $p < .001$. This finding replicates previous work demonstrating that expected linguistic material is easier to process than unexpected material.

Within the non-predictive conditions, pre-nominal modification clearly facilitated the processing of unexpected object nouns. Length-adjusted RTs for object nouns were faster for pre-nominal modification than for post-nominal modification, β = -124.86, SE = 30.42, $t(23.07) = -4.104$, $p < .001$. This result is consistent with the UID hypothesis, which predicts a processing advantage for the pre-nominal encoding: pre-modification reduces the surprisal of the unexpected word and results in a more uniform distribution of processing effort across the linguistic signal.

Within the predictive conditions, length-adjusted RTs for object nouns were also faster for pre-nominal modification than for post-nominal modification, β = -51.00, SE = 18.67, $t(29.26) = -2.73$, $p < .05$. However, the facilitation effect was weaker for predictive conditions, resulting in a context × encoding interaction, β = 73.67, SE = 33.02, $t(53.56) = 2.23$, $p < .05$. The UID hypothesis predicts a trough in information density for words that are both highly expected and pre-modified. Figure 2 (upper panel) is compatible with this prediction. RTs drop steeply in the predictive / pre-nominal condition at the object noun. Note that this is true despite the fact that object nouns were selected to be low-cloze. Thus, as shown in Figure 3, the post-nominal condition distributes the informational content more uniformly, resulting in a smoother RT profile.

Modification region analysis. Length-adjusted RTs for the modification region were analyzed using the same mixed effects model as above. Figure 3 (right panel) shows that encoding influenced the processing of the modification region in a way that was complementary to its effect on object nouns (see also Table 2). Pre-nominal modification was read more slowly than post-nominal modification in both contexts, β = 106.04, SE = 15.02, $t(75.75) = 7.06$, $p < .001$. However the magnitude of this effect was greater in the non-predictive context, reflected in a context × encoding interaction, β = -53.44, SE = 18.69, $t(51.68) = -2.86$, $p < .01$.

Table 2: Mean length-adjusted RT (ms) by condition for object noun (*upper panel*) and modification region (*lower panel*). Standard deviation in parentheses.

Object Noun	Pre-nominal	Post-nominal	*Mean*
Predictive	-54 (77)	-5 (54)	-30 (50)
Non-predictive	27 (70)	56 (98)	92 (46)
Mean	-13 (56)	76 (62)	
Modification Region	**Pre-nominal**	**Post-nominal**	***Mean***
Predictive	54 (46)	2 (39)	28 (31)
Non-predictive	113 (53)	10 (30)	61 (34)
Mean	84 (30)	6 (22)	

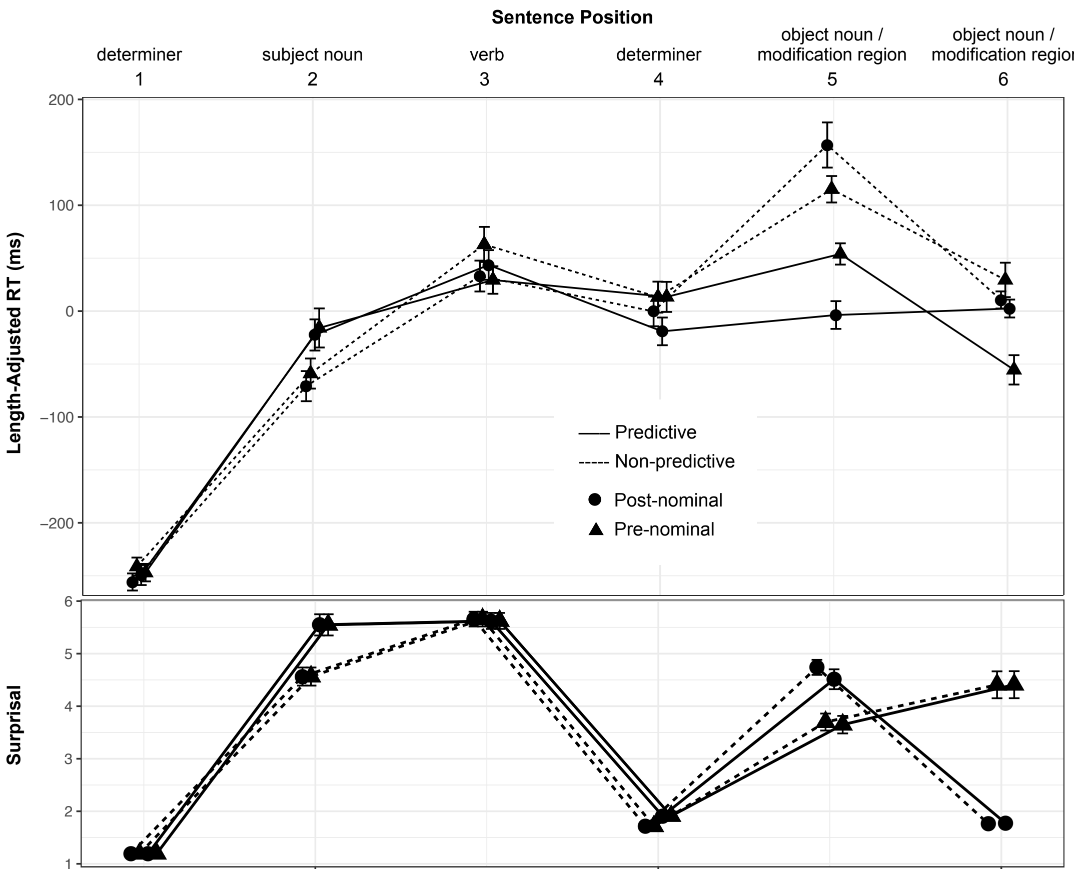

Figure 2: *Upper panel:* Mean length-adjusted word-by-word RTs. *Lower panel:* Surprisal profiles as determined by a language model trained on the German Wikipedia corpus. RTs and surprisals for the modification region were calculated by averaging across all modifiers words. Error bars represent one standard error of the mean.

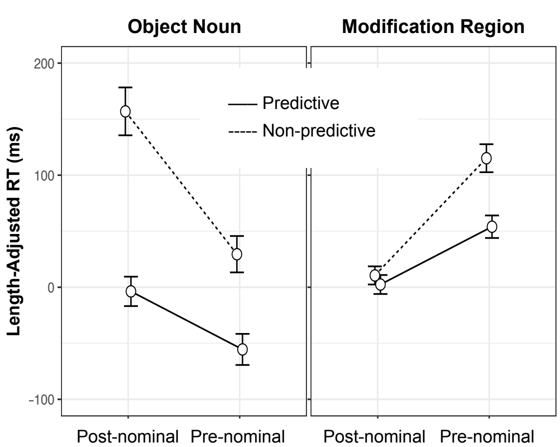

Figure 3: Mean length-adjusted RTs for object nouns (*left panel*) and the modification region (*right panel*). Error bars represent one standard error of the mean.

Surprisal profiles. The lower panel of Figure 2 shows the surprisal profiles produced by the language model. Surprisals at sentence positions 1-4 patterned closely with RTs, reflected in a positive correlation within this region (r = 0.36). However, the surprisal pattern differed somewhat from the pattern of RTs during the critical object NP region (r = -0.06). There are at least two plausible explanations for this divergence. First, the predictable contexts may not have made our atypical (i.e., low cloze) object nouns statistically more predictable, given our German Wikipedia corpus. For instance, *verpackte* ("boxed", a verb in the predictive context) and *Geschenk* ("gift", the corresponding object noun) were both present in the corpus but never co-occurred in the same sentence. We assessed this possibility and found that, on average, subject nouns in predictive contexts did not substantially increase the predictability of object nouns above the general case. Verbs, however, did so by a factor of 6. Second, the dependencies that existed in the training corpus may not have been fully captured by the language model. To test this possibility, we calculated the mean gram size used for surprisal predictions in the object NP region (M = 1.86). This finding indicates that despite being trained on 5-grams, the model predictions in this region were based predominantly on more local statistics (i.e., bigrams), effectively modeling only the non-predictive conditions.[7]

Despite these caveats, the results broadly confirm our assumptions about the distribution of surprisal across pre-nominal and post-nominal encodings of the critical object

[7] Note, however, that the cloze results validate both the stimuli and the RT findings.

NP: according to the language model, the pre-nominal encodings had more uniform information densities. To capture the behavior found for reading times in the predictive conditions, either a closer domain match between training corpus and stimuli would be required, or a language model architecture that is less sensitive to word position.

General Discussion

The UID hypothesis links production strategies with comprehension processes and predicts that speakers utilize flexibility in encoding to distribute information as evenly as possible across the linguistic signal (Jaeger, 2010; Levy & Jaeger, 2007). While prior evidence for UID comes primarily from word-level effects (Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007; Mahowald et al., 2013), a critical assumption underlying the UID hypothesis is that comprehenders should also be sensitive to the information density of alternative syntactically-complex encodings. To our knowledge, the current study is the first to investigate this important and challenging question.

We manipulated the syntactic encoding of complex noun phrases via meaning-preserving pre-nominal and post-nominal modification in contexts that were either predictive or non-predictive. The results were consistent with the UID hypothesis. In predictive contexts, post-nominal encodings elicited a more uniform distribution of processing effort than pre-nominal encodings. This makes sense because the head noun is already expected in such contexts, thus pre-nominal modification could lead to a trough in information density at the noun. Conversely, in non-predictive contexts, a more uniform RT profile was found for pre-nominal encodings, where pre-modification served to reduce the surprisal of the unexpected head noun. This pattern of comprehension results provides indirect support for UID as a rational strategy for producers to adopt. An important question for further investigation is whether speakers are indeed attentive to such factors when making their encoding decisions.

Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 *Information Density and Linguistic Encoding*.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes. 'R' package. Version 0.9975-12. http://CRAN. R-project.org

Drieghe, D., Rayner, K., & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance, 31*(5): 954–969.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*(3), 348-368.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods, 41*, 163-171.

Frank, A. & Jaeger, F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Conference of the Cognitive Science Society*, (Vol. 30, No. 30).

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,* (pp. 1-8). Association for Computational Linguistics.

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

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*: 262-284.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models (Version 2.0-30) http://cran.r-project.org/web/package= lmerTest

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

Levy, R. & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In Schöllkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, pp. 849-856, Cambridge, MA. MIT Press.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition, 126*(2): 313-318.

R Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org

Rayner, K., Aschby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the ez reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30*(12).

Rayner, K. & Well, A. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*(4): 504-509.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime: User's guide. Psychology Software Incorporated.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Interspeech* (Vol. 2002, p. 2002).

Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research, 41*(2), 105-128.