UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Prediction and uncertainty in an artificial language

Permalink

https://escholarship.org/uc/item/5dv7f9sw

Journal

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

Authors

Linzen, Tal
Siegelman, Noam
Bogaerts, Louisa

Publication Date

2017

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Prediction and uncertainty in an artificial language

Tal Linzen
LSCP & IJN, CNRS, ENS, PSL Research
University and Johns Hopkins University
tal.linzen@gmail.com

Noam Siegelman
Department of Psychology
Hebrew University of Jerusalem
noam.siegelman@gmail.com

Louisa Bogaerts
Department of Psychology
Hebrew University of Jerusalem
bog.louisa@gmail.com

Abstract

Probabilistic prediction is a central process in language comprehension. Properties of probability distributions over predictions are often difficult to study in natural language. To obtain precise control over these distributions, we created artificial languages consisting of sequences of shapes. The languages were constructed to vary the uncertainty of the probability distribution over predictions as well as the probability of the predicted item. Participants were exposed to the languages in a self-paced presentation paradigm, which provides a measure of processing difficulty at each element of a sequence. There was a robust pattern of graded predictability: shapes were processed faster the more predictable they were, as in natural language. Processing times were also affected by the uncertainty (entropy) over predictions at the point at which those predictions were made; this effect was less consistent, however.

Keywords: Entropy, prediction, statistical learning, artificial language, psycholinguistics

Introduction

Our environment is characterized by recurring temporal patterns; the sound of an ambulance siren, for example, tends to predict the appearance of an ambulance. Humans can quickly learn to exploit these contingencies between stimuli to anticipate future events and react to those events more effectively. The ability to track dependencies across the elements of a sequence is central to language processing: prediction of upcoming words is employed during language comprehension (DeLong, Urbach, & Kutas, 2005) and may play a central role in acquisition (Gómez, 2002).

Prediction in natural language is rarely categorical: there is generally some uncertainty as to the upcoming word. Rather than predict a single word or avoid making predictions altogether, readers maintain a probability distribution over the upcoming words: words that are more likely to come up are activated to a greater extent (Smith & Levy, 2013). Probability distributions over predictions are often difficult to study in natural language, due to the need to find sets of words that happen to have the desired probabilistic relations in a natural corpus. The present study builds on work that shows that the processing of temporal contingencies can be studied using artificial language learning experiments. These experiments typically consist of a familiarization phase, in which participants are exposed to the artificial language, and a test phase, in which they are requested to distinguish sequences that follow the patterns of the language from sequences that do not. We use this paradigm to study probabilistic prediction in sequence learning and processing.

Quantifying probabilistic prediction: A predictive dependency is made up of two parts: the point at which the prediction is generated (the predictive item) and the point at which it is matched against the incoming input (the predicted item). We study both parts of the process. At the predictive item, multiple probabilistic predictions can typically be generated. Higher uncertainty over the correct prediction may lead to increased competition among those predictions and slower processing. We follow earlier work in quantifying uncertainty using the *entropy* of the distribution over possible predicted items (Linzen & Jaeger, 2014; Hasson, 2017):

$$H = -\sum_{w \in W} P(w) \log_2 P(w) \tag{1}$$

where W is the set of possible items and $P(w)$ is the probability of w in the current context. At the point at which predictions are matched against the input, input items that were predicted with a higher probability may be processed more quickly. In natural language, processing difficulty at an item w is proportional to its *surprisal* ($-\log_2 P(w)$): more surprising words tend to be read more slowly (Smith & Levy, 2013). A final expectation-based measure that has been argued to be a reliable predictor of reading times (RTs) in natural language is *uncertainty reduction*: words that reduce uncertainty about the sequence to a greater extent are predicted to be read more slowly (Hale, 2003; Frank, 2013).

The experiments: We report two experiments designed to examine these quantitative measures of probabilistic prediction in artificial languages. In what follows, we briefly discuss our general methodological strategy.

In many artificial language learning experiments, the familiarization phase consists of passive exposure; as such, the only behavioral measure collected in these studies is the proportion of correct grammaticality judgments given after the familiarization stage is over. Recently, a number of online paradigms have been proposed that track the learning process as it unfolds over the course of the familiarization phase (Siegelman, Bogaerts, Christiansen, & Frost, 2017). Online paradigms also provide an index of processing time at each individual item of the sequence, making them particularly well-suited to studying the generation and validation of predictions. We adopt one of these paradigms, the self-paced reading paradigm (Just, Carpenter, & Woolley, 1982), adapted to artificial language learning by Karuza, Farmer, Fine, Smith, and Jaeger (2014). In this paradigm, the elements of each sequence are presented sequentially; the participant controls when the next sequence element is revealed.

Previous studies of prediction have focused on sequences with nonadjacent dependencies: the sequence is of the form

AXB, where A predicts B (Karuza et al., 2014; Misyak, Christiansen, & Bruce Tomblin, 2010). Instead, we use dependencies of the form AB, without an intervening element; such dependencies are in general easier to learn (Newport & Aslin, 2004). By increasing the likelihood that our participants will learn the language, we can ask more fine-grained questions than would be possible using nonadjacent dependencies.

Summary of goals: We address the following issues:

1. Does the probability of the second shape B given the first shape A affect the processing of B?
2. Are processing times at the point where the prediction can be made (shape A) affected by the uncertainty of the probability distribution over predictions?
3. Does the reduction of uncertainty about the sequence at shape B entail greater processing difficulty?

Experiment 1

Stimuli

Following Karuza et al. (2014), we used sequences of letters from the Ge'ez script, which is used to write several Ethiopian and Eritrean languages. We refer to these letters as shapes since none of our participants were familiar with this script. As in Karuza et al. (2014), our sequences consisted of three shapes. As we have mentioned, we omitted the intermediate shape – the dependency was adjacent. To keep the structure of the stimuli similar to the stimuli used in the previous study and to avoid task effects related to the beginning of a new sequence, all of the sequences started with a fixed shape r (distinct from the A and B shapes). This shape was the same in all trials for a given participant and was not analyzed.

The language used in this experiment is described in Table 1. Each participant was exposed to two types of A shapes. Low entropy A shapes were followed by one of two B shapes, with probability $1/4$ and $3/4$ respectively. High entropy A shapes were followed by one of four B shapes, each with probability $1/4$. There were two A shapes of each type, for a total of four A shapes. None of the B shapes were repeated across A shapes: there were 12 distinct B shapes. We refer to the B shapes with a probability of $1/4$ as high surprisal shapes, and to shapes with a probability of $3/4$ as low surprisal shapes.

To control for potential differences in the visual complexity of particular shapes, the shapes that served as a_1, a_2 and $b_1, \ldots, b_6$ were counterbalanced across participants.

Participants

A total of 44 participants (24 women and 20 men; age range: 20–28, mean age: 23.4) from the Hebrew University of Jerusalem community completed the experiments.

Procedure

The experiment consisted of three phases: familiarization, test and a post-test phase.

Shape R	Shape A	Shape B	TP	Surprisal
High entropy: ($H = 2$)				
r	a_1	b_1	$1/4$	2
r	a_1	b_2	$1/4$	2
r	a_1	b_3	$1/4$	2
r	a_1	b_4	$1/4$	2
Low entropy: ($H = 0.81$)				
r	a_2	b_5	$1/4$	2
r	a_2	b_6	$3/4$	0.41

Table 1: Half of the language used in Experiment 1 (the other half is duplicated: a high entropy a_3 paired with high surprisal b_7 through b_{10} and a low entropy a_4 paired with a high surprisal b_{11} and a low surprisal b_{12}). TP indicates the transitional probability between the A and the B shape (e.g., $P(b_1|a_1) = 1/4$). H indicates the entropy of each distribution.

Familiarization phase: Each trial started with a sequence of dashes where the shapes would be; participants pressed the spacebar to reveal the next shapes one by one. When a shape was revealed, the previous shape was replaced by a dash again. Before this phase of the experiment began, participants were instructed to try to remember the sequences, since they would be tested on them later on.

There were 288 sequences in this phase. This contrasts with the familiarization phase in Karuza et al. (2014), which consisted of 432 sequences; we chose to have a shorter familiarization phase because prediction effects in Karuza et al. (2014) plateaued about half way through the experiment. We further simplified their design by eliminating the catch trials meant to ensure that participants were paying attention. These trials were not necessary because we analyzed data only from participants who successfully learned the language: our assumption was that participants who were not paying attention would fail to learn the language.

Test phase: This phase consisted of 24 trials, each of which elicited a judgment for one sequence. All three shapes of the sequence were presented at once (not in self-paced presentation). Half of the trials contained sequences that had been presented during familiarization; the other half contained the shapes from the familiarization phase arranged in unseen sequences. Participants were asked to press one button if the sequence appeared familiar given the sequences they had seen in the first phase, and another button if it did not.

Post-test phase: The test phase was followed by another self-paced presentation phase. This phase was somewhat shorter, consisting of 192 trials. Participants were again instructed to attempt to remember the shapes. The goal of this phase was to examine the behavior of participants who have already learned the language; for example, if predictability effects were found, are they restricted to the stages in which the participant has not yet mastered the language?

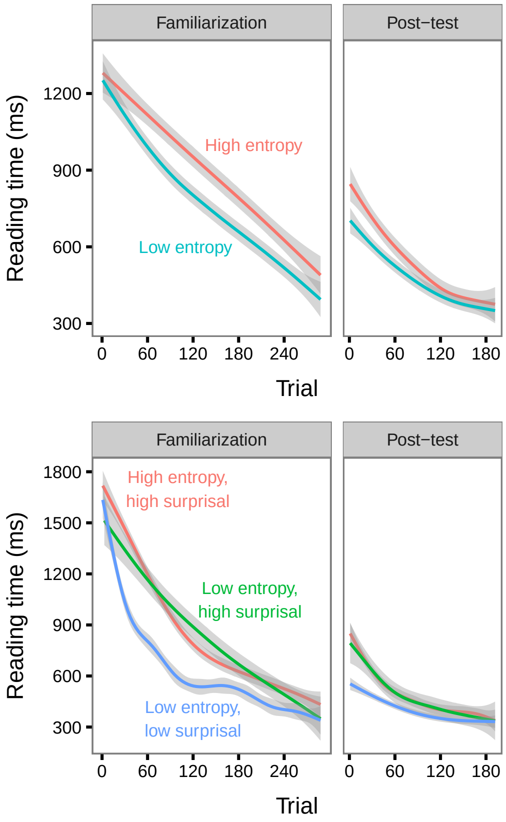

Figure 1: Reading times in Experiment 1 (above: *A* shape; below: *B* shape).

Results

Accuracy: We briefly analyze the familiarity judgments from the test phase before moving on to the analysis of the processing time data from the familiarization phase, which is the focus of this study. On average, participants were more likely to judge a sequence as grammatical, leading to higher accuracy on grammatical than ungrammatical sequences (82% vs. 60%). To test for differences across types of test sequences, we coded the test sequences based on the category of their *A* and *B* shapes (e.g., low entropy + high surprisal). Logistic mixed-effects models fitted separately to grammatical and ungrammatical sequences did not find significant differences across sequence types (grammatical: $\chi^2(2) = 2.8$, $p = .24$; ungrammatical: $\chi^2(2) = 4.4$, $p = .11$).

RT preprocessing and analysis: We refer to the sequential processing times measured by key press latencies as reading times (RTs) for consistency with the sentence processing literature. Following Karuza et al. (2014), we excluded shapes on which RTs were (1) longer than six seconds or (2) three standard deviations higher or lower than the participant's mean RT for shapes in the same position. This resulted in the exclusion of 1.1% and 2.5% of the shapes respectively.

We only analyzed RTs from participants who gave correct grammaticality judgments to at least 18 of the 24 sequences (the lowest number for which $p < .05$ according to an exact binomial test). Of the 44 participants, 23 passed this threshold. Our statistical analysis largely followed Karuza et al. (2014). RTs were log-transformed and submitted to a linear mixed-effects regression with a random intercept for shape and a random intercept and slope for all fixed effects. Trial number and its interaction with the experimental factors were included in all models.

RT results: The time course of the results is plotted in Figure 1. Overall, RTs decreased markedly over the course of the familiarization phase, picked up in the beginning of the post-test phase, then decreased again.

The average difference in RTs between high and low entropy *A* shapes in the familiarization phase was 119 ms (877 ms for high entropy and 758 ms for low entropy shapes). The linear mixed-effects model analysis indicated that this difference was statistically significant ($\chi^2(1) = 4.2$, $p = .04$). The effect of trial number was highly significant ($\chi^2(1) = 31.5$, $p < .001$). The interaction between trial number and entropy did not reach significance ($\chi^2(1) = .06$, $p = .81$), suggesting that there was no clear evidence that the effect of entropy changed over the course of the experiment.

There were three types of *B* shapes: high surprisal ones that followed a low entropy *A* shape (e.g., b_5, see Table 1); high surprisal ones that followed a high entropy *A* shape (e.g., b_1); and low surprisal ones that followed a low entropy *A* shape (e.g., b_6). We first examined the effect of surprisal, collapsing across the two categories of high surprisal shapes. We found that high surprisal shapes were read more slowly than low surprisal shapes ($\chi^2(1) = 17.8$, $p < .001$); the average difference in RT was 200 ms (812 ms for high surprisal and 612 ms for low surprisal shapes). The effect of trial number was highly significant again ($\chi^2(1) = 44.6$, $p < .001$), and interacted with surprisal such that the effect of surprisal weakened over the course of the familiarization phase ($\chi^2(1) = 9.9$, $p = .002$).

Finally, we compared the two types of high-surprisal *B* shapes, which were matched for surprisal but differed in the entropy of the *A* shape that preceded them. The mean RTs were almost identical across these two types of shapes (812 ms after high entropy *A* shapes and 813 ms after low entropy ones). This difference was not significant in the statistical analysis (main effect of entropy: $\chi^2(1) = 1.2$, $p = .27$; interaction with trial number: $\chi^2(1) = .9$, $p = .35$).

Discussion

In this experiment, participants were taught a language designed to assess the effect of measures of probabilistic prediction on sequence processing. Neither surprisal nor uncertainty reliably affected judgment accuracy in the test phase; they did, however, modulate processing times during the familiarization phase. First, predictability affected RTs in the expected way: high surprisal *B* shapes were read more slowly than low surprisal ones. Second, uncertainty at the *A* shape affected RTs in a way that is consistent with competition among the predictions: higher entropy shapes were read more slowly than low entropy ones.

Finally, we did not find evidence for an effect of uncertainty reduction on the *B* shape. To see why, note that the *B* shapes are the last item in the sequence; as such, they reduce the uncertainty about the sequence to 0. The amount by which uncertainty is reduced is therefore equal to the entropy of the distribution over predictions at the *A* shape; yet there was no evidence for a difference in reading times between high surprisal *B* shapes that followed a high entropy *A* shape (and therefore reduced entropy by 2 bits) and high surprisal *B* shapes that followed a low entropy *A* shape (and reduced entropy by only 0.41 bits).

Experiment 2

In Experiment 1, uncertainty was perfectly correlated with the number of possible predictions: high entropy *A* shapes had four prediction options compared to two options in low entropy *A* shapes. The goal of the current experiment is to examine whether we can find entropy effects when the number of options is kept constant. For a given number of options, entropy is highest when the distribution is uniform; we therefore compare a uniform distribution to a skewed one, that is, with one option that is more likely than the others.

Participants

A total of 49 participants completed the experiment. Two participants were excluded for not completing the experiment and one for having prior exposure to Amharic, which uses the Ge'ez script; of the remaining participants, 35 were women and 11 men (age range: 19–31; mean age: 23.8).

Materials

The language used in Experiment 2 is shown in Table 2. There were three types of *A* shapes. Two of the *A* shapes could be followed by three possible *B* shapes (to avoid having to teaching participants a very low probability option, we used three options instead of four as in Experiment 1.) After a_1, the distribution of the *B* shapes was uniform: each of the shapes had a probability of $1/3$. After a_2 the distribution was skewed: one of the shapes had a probability of $2/3$ and the other two $1/6$ each.

To control for the possibility that any difference between the two 3-option shapes could reflect skew rather than entropy as such, we additionally included a third type of *A* shape that was followed by one of *two B* shapes, each with probability $1/2$. As this distribution is uniform, we expect this shape to pattern with a_1 if the relevant factor is skew. Conversely, since its entropy is lower than either 3-option shapes, it should be processed faster than either of them if entropy is the relevant factor.

Due to the larger number of conditions and the need to provide sufficient exposure to lower probability *B* shapes ($1/6$ compared to $1/4$ in Experiment 1), each type of *A* shape was represented by a single shape only.

Shape 1	Shape 2	Shape 3	TP	Surprisal
Three options, uniform: ($H = 1.58$)				
r	a_1	b_1	$2/6$	1.58
r	a_1	b_2	$2/6$	1.58
r	a_1	b_3	$2/6$	1.58
Skewed, three options: ($H = 1.25$)				
r	a_2	b_4	$4/6$	0.58
r	a_2	b_5	$1/6$	2.58
r	a_2	b_6	$1/6$	2.58
Uniform, two options: ($H = 1$)				
r	a_3	b_7	$3/6$	1
r	a_3	b_8	$3/6$	1

Table 2: Language used in Experiment 2. *H* indicates the entropy of each distribution.

Procedure

The structure of the experiment was the same as in Experiment 1. The familiarization self-paced presentation phase consisted of 324 sequences. This phase was followed by 16 familiarity judgments, and an additional post-test self-paced presentation phase with 216 sequences.

Results

Accuracy: Overall accuracy was higher than in Experiment 1, though the bias for marking sequences as grammatical remained: 93% of the grammatical sequences and of 77% of the ungrammatical sequences were identified correctly. We tested for an effect of the four types of sequences (see Table 2) on accuracy rates on grammatical sequences. A logistic mixed-effects model did not reveal an effect of sequence type ($\chi^2(3) = 4.5$, $p = .21$). Likewise, there was no effect of either *A* or *B* shape type on accuracy rates in ungrammatical sentences (*A*: $\chi^2(2) = 2.65$, $p = .27$; *B*: $\chi^2(3) = 3.2$, $p = .36$).

RT preprocessing and analysis: As before, we restricted our analysis to participants who showed evidence of learning the language, defined as giving correct judgments more often than chance ($p < .05$ according to the binomial test); this translates to performing at least 13 of the 16 trials correctly. Of the 46 participants, 33 passed this threshold. We excluded key presses with extreme RTs using the same criteria as before, resulting in the exclusion of 3.38% of the shapes. Analysis methods were in general identical to Experiment 1, with the exception that our mixed-effects models did not include a random intercept for shape in cases where there was only one shape in each condition (i.e., in the analysis of *A* shapes).

RT results: The qualitative pattern of results was similar to Experiment 1: RTs globally decreased over the course of the familiarization phase, briefly increased in the post-test phase, then decreased again.

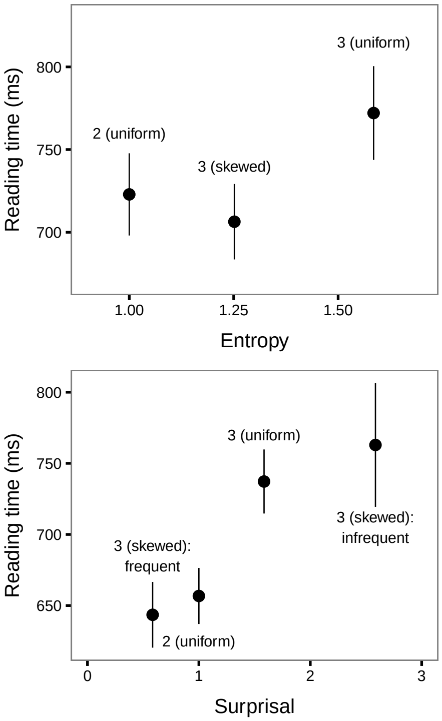

Figure 2: Condition means in the familiarization phase of Experiment 2 (above: *A* shape; below: *B* shape). Error bars represent 95% within-subject confidence intervals.

We first discuss the statistical analysis of familiarization phase RTs on *A* shapes, starting with an analysis of entropy as a numerical predictor There was a main effect of entropy ($\chi^2(1) = 5.1$, $p = .02$), a main effect of trial number ($\chi^2(1) = 40.3$, $p < .001$) and a nonsignificant interaction ($\chi^2(1) = 3.3$, $p = .07$). RTs in the individual conditions were longest on the uniform 3-option shape and shortest on the skewed 3-option shape; although the entropy of the 2-option shape was lowest of all three shapes, average reading times on this shape were somewhat higher than the skewed 3-option shape (see Figure 2). The difference in RTs between the two 3-option shapes was significant ($\chi^2(1) = 5.1$, $p = .02$), but the interaction with trial number was not ($\chi^2(1) = 2.3$, $p = .13$). The difference between the two shapes with a uniform prediction distribution (3-option vs. 2-option) and the interaction between this difference and trial number did not reach significance (main effect: $\chi^2(1) = 3.5$, $p = .06$; interaction: $\chi^2(1) = 3.1$, $p = .08$), and neither did the difference between the skewed 3-option and uniform 2-option shapes (main effect: $\chi^2(1) = .5$, $p = .48$; interaction: $\chi^2(1) = .03$, $p = .87$).

We next discuss the *B* shapes. Again, we first enter surprisal as a numerical predictor. The statistical analysis found a highly significant effect of this predictor ($\chi^2(1) = 36.2$, $p < .001$) and of trial number ($\chi^2(1) = 75.3$, $p < .001$), as well as an interaction between the two ($\chi^2(1) = 20.3$, $p < .001$). Inspection of the average RTs for each level of surprisal (see Figure 2) suggests that not all differences between consecutive levels of surprisal are equally large; in fact, only the difference between the $p = 2/6$ and $p = 3/6$ shapes was statistically significant ($\chi^2(1) = 21.3$, $p < .001$).

Discussion

RTs on the two 3-option *A* shapes were consistent with the hypothesis that higher uncertainty leads to longer RTs. The difference was smaller than in Experiment 1 (around 60 ms), though that is to be expected given the smaller difference in entropy between the two shapes in the current experiment. The same hypothesis, however, predicts that RTs on the 2-option shape should be lower than either 3-option shape; there was no evidence for such an effect.

There was a strong effect of surprisal overall, but there was often no evidence for differences between consecutive levels of surprisal. The difference between the two *B* shapes that followed the 3-option skewed *A* shape was particularly large. Finally, since no two *B* shapes were matched on predictability and at the same time differed in the entropy of the *A* shape that predicted them, the design of Experiment 2 did not allow us to test for an effect of uncertainty reduction.

General Discussion

Probabilistic prediction plays a central role in language processing: a predictive item sets up expectations for predicted items later in the sequence. We studied the reflexes of probabilistic prediction in two artificial languages, which allowed us to exert precise control over the distribution over predictions. We used self-paced presentation (Just et al., 1982; Karuza et al., 2014), which yields implicit measures of processing at every element of the sequence. Two experiments revealed graded predictability effects parallel to those found in natural language. They also suggested that higher uncertainty over predictions at the point where predictions are generated leads to longer processing times, although these effects were weaker. No clear support was found for an effect of uncertainty reduction, even when controlling for predictability.

To further investigate the results, we pooled the data from both experiments and plotted the mean RTs in the familiarization phase by numerical entropy and surprisal in Figure 3 (since Experiment 2 was slightly longer, we discarded the trials following the first 288 trials for the purpose of this analysis). The evidence for a linear effect across experiments of the numerical predictors appears stronger for surprisal than for entropy. In particular, there are no clear differences among low-entropy distributions (lower than 1.5), and the slight differences that do exist are in the opposite direction than predicted by a linear relationship between entropy and RTs. Statistical models including data from both experiments did not reveal overall entropy effects at the *A* shape (entropy: $\chi^2(1) = 0.4$, $p = .53$; trial number: $\chi^2(1) = 63$, $p < .001$; interaction: $\chi^2(1) = 2$, $p = .16$), but did reveal clear surprisal effects at the *B* shape as well as an interaction with trial number (surprisal: $\chi^2(1) = 25.5$, $p < .001$; trial number: $\chi^2(1) = 104.1$, $p < .001$; interaction: $\chi^2(1) = 21.7$, $p < .001$).

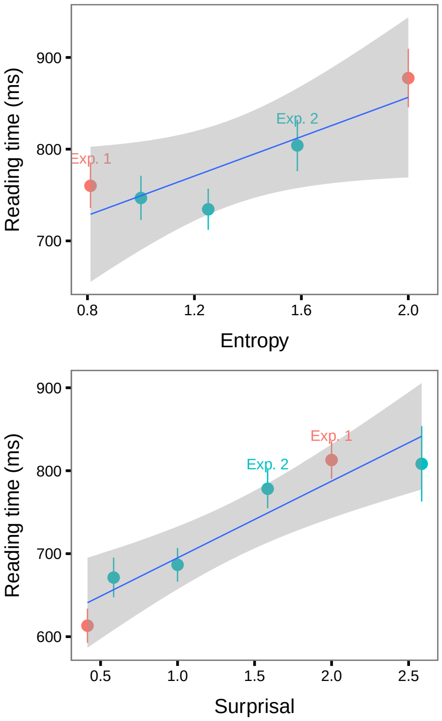

Figure 3: Comparison across the experiments: means of the first 288 trials of the familiarization phase (above: *A* shape; below: *B* shape). Error bars represent within-subject confidence 95% confidence intervals based on two standard deviations from the mean.

While any conclusion from pooling together two experiments with a different design and a different set of subjects should be taken as tentative, the nonlinear relationship between entropy and processing times suggests that entropy may not be the best metric for difficulty in prediction generation; additional properties of the distribution over predictions, such as the number of options or the probability of the most likely option, may need to be taken into consideration.

Figure 1 suggests that RTs in Experiment 1 may have reached a plateau about 250 trials into the familiarization phase; differences among conditions appeared to grow increasingly small around this time (Karuza et al. (2014) report a similar pattern). RTs increased at the beginning of the post-test phase, and then plateaued again around 100 trials into the pre-test phase. We did not present an in-depth analysis of the post-test phase for reasons of space; however, the fact that the overall increase in RTs at the beginning of the post-test phase was accompanied by a re-emergence of predictability and entropy effects suggests that the convergence between the conditions at the end of the familiarization phase is due to a floor effect rather than due to participants abandoning predictive processes once the language has been learned.

We made relatively few modifications to the methodology developed by Karuza et al. (2014), with the goal of building on their established paradigm. This entailed in particular that our sequences were made up of visual symbols rather than auditory or written words; none of the symbols had any semantic content. The encouraging results of the present study suggest that this method may be extended to richer artificial languages that are a closer approximation of natural languages.

Acknowledgments: We thank Lizz Karuza for sharing her materials. This research was supported by the European Research Council (grant ERC-2011-AdG 295810 BOOTPHON) and the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC).

References

DeLong, K., Urbach, T., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*(3), 475–494.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Hasson, U. (2017). The neurobiology of uncertainty: implications for statistical learning. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160048.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.

Karuza, E. A., Farmer, T. A., Fine, A. B., Smith, F. X., & Jaeger, T. F. (2014). On-line measures of prediction in a self-paced statistical learning task. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 725–730).

Linzen, T., & Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. In *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18).

Misyak, J. B., Christiansen, M. H., & Bruce Tomblin, J. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*(1), 138–153.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I: Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosphical Transactions of the Royal Society B*, *372*(1711), 20160059.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.