Uncertainty and expectation in sentence processing

Tal Linzen¹ and T. Florian Jaeger²

¹New York University and ²University of Rochester

Author Note

Tal Linzen, Department of Linguistics, New York University.


A previous version of this work was presented at the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics. Acknowledgements withheld for the review process. This work was supported by an Alfred P. Sloan Fellowship and NSF CAREER (IIS-1150028) to T. Florian Jaeger.

Correspondence concerning this article should be addressed to Tal Linzen, Department of Linguistics, New York University, New York, NY 11237.

Contact: linzen@nyu.edu
Abstract

There is now considerable evidence that human sentence processing is expectation based: As people read a sentence, they use their statistical experience with their language to generate predictions about upcoming syntactic structure. This study examines how sentence processing is affected by readers’ *uncertainty* in their expectations about upcoming structure. In a self-paced reading study, we manipulate both expectations and the uncertainty about them factorially, based on lexical subcategorization information. We compare two types of uncertainty: uncertainty about the verb’s complement, reflecting the next prediction step; and uncertainty about the full sentence, reflecting an unbounded number of prediction steps. We find that uncertainty about the full structure, but not about the next step, was a significant predictor of processing difficulty: Greater reduction in uncertainty was correlated with increase reading times. We additionally replicated previously observed effects of expectation violation (surprisal), orthogonal to the effect of uncertainty. This suggests that both surprisal and uncertainty have a role to play in explaining human reading times. We discuss the consequences for theories of sentence comprehension.

*Keywords:* Sentence processing, uncertainty, prediction, entropy reduction, surprisal, competition
Uncertainty and expectation in sentence processing

One of the major challenges that readers face when processing a sentence is inferring its syntactic structure (parsing it). There is growing evidence that people parse sentences in an incremental and predictive fashion: Each incoming word is used to revise existing hypotheses about the correct parse of the sentence and predict upcoming syntactic structure (Altmann & Kamide, 1999; Hale, 2001; Levy, 2008). These predictions are probabilistic: There is a continuous relationship between predictability and processing difficulty (Boston, Hale, Kliegl, Patil, & Vasishth, 2010; Demberg & Keller, 2008; Jennings, Randall, & Tyler, 1997; McDonald & Shillcock, 2003a, 2003b). This likely reflects a strategy whereby readers prepare to process upcoming linguistic material in proportion to its probability (DeLong, Urbach, & Kutas, 2005; Smith & Levy, 2013).

As an example, consider the sentence in (1):

(1) He accepted the proposal was wrong.

The verb accept can be followed by two types of complements, or subcategorization frames: a noun phrase (as in accept a gift) or a sentential complement (as in accept that you’ve lost). In actual usage, accept occurs much more frequently with the noun phrase (NP) frame than with the sentential complement (SC) frame. Having read the word accepted, then, the reader can form a strong prediction for an NP, and possibly a weaker prediction for an SC (Fig. 1a and 1b). The next words, the proposal, are compatible with both parses: They can either serve as the verb’s direct object or as the subject of an SC (Fig. 1c and 1d). Finally, the words was wrong disambiguate the sentence in favor of the low-probability SC parse. In line with the predictive parsing hypothesis, the disambiguating region was wrong tends to be read more slowly when the
verb favors the NP frame (e.g., *accept*) than when it favors the SC frame (e.g., *prove*) (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell, Tanenhaus, & Kelly, 1993).

While there is ample evidence that expectation violation, as in the example just discussed, can lead to processing difficulty, less is known about the generation and maintenance of those expectations. Suppose, for example, that the verb *accepted* in sentence (1) were replaced by *forgot*. The verb *forgot* is similar to *accept* in that it is biased against an SC continuation, but differs from it in that it has a more diverse set of potential complements. Specifically, in addition to an NP (*forgot my birthday, 55%*) and an SC (*forgot he was supposed to go, 9%*), this verb can be followed by a prepositional phrase (*forgot about the party, 18%*) or an infinitive (*forgot to buy groceries, 14%*). Consequently, there is a greater degree of uncertainty about upcoming syntactic structure after *forget* than after *accept*. Does this difference between *forget* and *accept* affect processing difficulty, and if so, how?

Following standard practice, we quantify uncertainty about a probabilistic outcome using the Shannon entropy of the distribution:

\[
H = -\sum_{i=1}^{n} p_i \log_2 p_i
\]

Entropy is higher the more potential parse completions there are and the more distributed the probabilities of those parses are. For instance, when there is only one parse completion, the entropy is 0 bits. With two equiprobable completions the entropy is 1 bit, and with four equiprobable options the entropy is 2 bits. However, if one of the four options is much more likely than the others, the entropy can be lower than the entropy of two equiprobable options (see Fig. 2).
We focus on three ways in which readers’ uncertainty about syntactic expectations, as quantified by entropy, may affect processing difficulty. First, it may be costly to generate and maintain a larger number of predictions that compete with each other, especially if their probabilities are similar. We term this hypothesis the *competition hypothesis* (Elman, Hare, & McRae, 2005; McRae, Spivey-Knowlton, & Tanenhaus, 1998). A second hypothesis, the *entropy reduction hypothesis*, proposes that it is reduction in uncertainty that is costly rather than the mere existence of uncertainty (Hale, 2006). Under this hypothesis, an increase in uncertainty does not affect processing. Finally, uncertainty over expectations may not affect reading times at all if expectation-related processing cost arises only when the expectation is matched against incoming words, as argued in surprisal theory (Hale, 2001; Levy, 2008).

The examples discussed so far have focused on the expectations that comprehenders generate immediately after processing a verb based on its subcategorization frequencies – in other words, expectations for the next immediate node in the parse tree following the verb (Fig. 1). Yet it is possible that readers generate more detailed syntactic predictions, consisting of multiple derivation steps. For instance, instead of simply predicting an NP, they might probabilistically predict both an NP consisting of a determiner and a noun (*the present*) and an NP consisting of a determiner, an adjective and a noun (*the nice present*) (Fig. 3). The depth of syntactic structure predicted by readers is an open question. We investigate two endpoints of the prediction depth spectrum – prediction of the next syntactic step of the derivation (*single-step prediction*) and prediction of the entire syntactic structure of the sentence (*full prediction*) – while acknowledging that in practice readers may predict variable amounts of syntactic structure, depending on a range of factors.
A small number of recent studies have attempted to explore the effect of uncertainty on reading times. One class of studies has demonstrated that the entropy reduction hypothesis can predict qualitative differences in processing difficulty across constructions, such as the asymmetry between object and subject relative clauses (Hale, 2003; Yun, Whitman, & Hale, 2010). These predictions, however, have not been evaluated on reading times, and have not controlled for alternative sources for the difference in processing difficulty between the constructions, such as surprisal (Hale, 2001), memory cost (Grodner & Gibson, 2005) or similarity based interference (Lewis & Vasishth, 2005).

Other studies have assessed the effect of entropy in reading times corpora (Frank, 2013; Roark, Bachrach, Cardenas, & Palier, 2009; Wu, Bachrach, Cardenas, & Schuler, 2010). These studies have produced a mixed pattern of results: Roark et al. (2009) found a positive effect of single-step entropy, supporting the competition hypothesis; Wu et al. (2010) found that entropy reduction had a positive effect on reading times, but only for closed class words; and Frank (2013) found a positive effect of entropy reduction, for all words. These mixed results are hard to evaluate for several reasons. First, the studies differed in the depth of prediction that underpinned the entropy estimates they used (single-step in Roark et al., 2009, four steps in Frank, 2013, full in Wu et al., 2010), and it is unclear how depth of prediction affects entropy estimates. Second, these studies have used a wide variety of syntactic frameworks, ranging from a connectionist network (Frank, 2013), through a lexicalized and heavily transformed Probabilistic Context-Free Grammar (Roark 2001, Roark et al., 2009) to a Hierarchical Hidden Markov Model (Wu et al., 2010). Third, one of these studies only measured effects of entropy (Roark et al., 2009) while the others only examined entropy reduction (Frank, 2013; Wu et al.,
Finally, all of these studies examined processing difficulty in reading time corpora, across different types of structures. While this approach evaluates the breadth of coverage of the tested hypotheses, it again introduces multiple potential confounds; none of the studies controlled for memory cost, for instance.

The current study attempts to clarify this picture by combining the advantages of the two approaches. Following Roark et al. (2009), we evaluate the predictions of uncertainty-related hypotheses on human reading times. Following Hale (2003), we undertake a detailed experimental and computational analysis of a specific class of sentences, using a simple syntactic framework: a probabilistic context-free grammar (PCFG) based on the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993). We emphasize two distinctions that are not always clearly highlighted in the literature: first, we distinguish entropy effects, which may support the competition hypothesis, from entropy reduction effects; and second, we compare uncertainty in single-step prediction and uncertainty in full prediction, and show that the two may differ dramatically. To avoid the potential confounds associated with comparing reading times across constructions, we manipulate syntactic uncertainty while keeping syntactic structure (and thus confounding factors such as memory cost) constant. We do so by varying the syntactic expectations induced by specific lexical items, as in the case of accept compared to forget above.

We first describe a reading time experiment in which we vary single-step entropy by comparing verbs with different subcategorization distributions. We then discuss the relationship between single-step and full entropy, and assess how well each of these measures predict reading times.

**Reading times experiment**
The design of the experiment was modeled after Garnsey et al. (1997). Half of the sentences read by a given participant included the complementizer *that*, as in (2a) (henceforth referred to as *unambiguous sentences*), and half did not, as in (2b) (henceforth *ambiguous sentences*).

(2) a. The men discovered that the island had been invaded by the enemy.
   
   subject verb that ambiguous disambiguating rest

   b. The men discovered the island had been invaded by the enemy
   
   subject verb ambiguous disambiguating rest

In addition to the within-item ambiguity manipulation, two factors were manipulated between items: the subcategorization entropy of the main verb (high vs. low), and the surprisal of an SC given the verb (high vs. low). Subcategorization frequencies were taken from Gahl, Jurafsky and Roland’s (2004) database (described in more detail below). We quantified SC bias using the surprisal (inverse log probability) of an SC given the verb rather than raw conditional probability, based on evidence that the relationship between conditional probability and processing difficulty is logarithmic (Smith & Levy, 2013). In summary, the reading time study used a 2 x 2 x 2 design: presence or absence of a complementizer x subcategorization entropy x SC surprisal.

If increased uncertainty causes a processing slowdown, as argued by the competition hypothesis, reading times at the verb region will be longer in the high subcategorization entropy conditions. Conversely, if it is reduction in uncertainty that causes processing slowdown, as argued by the entropy reduction hypothesis, we expect longer reading times after verbs with lower subcategorization entropy: Since entropy before the verb is identical across conditions (see
below), verbs with lower subcategorization entropy reduce uncertainty more than verbs with higher subcategorization entropy. Finally, the surprisal hypothesis predicts no effect of subcategorization entropy. It does predict that disambiguation in favor of an SC parse will be more costly for high SC surprisal verbs than for low SC surprisal verbs (Garnsey et al., 1997). This should only occur in ambiguous sentences.

**Method**

**Participants.** A total of 128 participants were recruited through Amazon Mechanical Turk and were paid $1.75 for their participation.

**Materials.** We selected 32 verbs, eight in each of the cells of the 2 x 2 design defined by subcategorization entropy and SC surprisal. Subcategorization frequencies were obtained from the database of Gahl et al. (2004), which is based on the 18 million words of text comprising the Touchstone Applied Science Associates corpus (Zeno, Ivens, Millard, & Duvvuri, 1995) and the Brown corpus (Francis & Kučera, 1982). Verbs were matched across conditions for their frequency and length. Frequency norms were obtained from the SUBTLEX-US corpus (Brysbaert & New, 2009). Table 1 shows the mean values and standard deviations across conditions for log-transformed verb frequency, subcategorization entropy and SC surprisal.²

INSERT TABLE 1 HERE.

In the next step, 32 sentences pairs were created, one for each verb (a list of all items is provided in Appendix A). Each pair contained one version of the sentence with the complementizer *that* after the verb and one without it (64 sentences in total). The matrix subjects of the sentences were chosen to be minimally informative two-word noun phrases (e.g. *the men*, *they all*), to avoid biasing the distribution over verb complement frames ahead of the verb. The same eight matrix subjects were used in all four conditions.³ Following the complementizer (or
the verb, if the complementizer was omitted) was a definite noun phrase (*the island*), which was always a plausible direct object of the verb (following Garnsey et al., 1997). The frequency of the noun was matched across conditions.

The disambiguating region consisted of three words: either two auxiliary verbs (*had been*) or an auxiliary verb and negation (*might not*), followed by the past participle of a verb (*invaded*). Each of the function words appeared the same number of times in each condition. The verbs (*invaded*) were matched across conditions for frequency and length. The disambiguating region was followed by three more words, which were not analyzed.

In addition to the target sentences, the experiment included 64 filler sentences. These sentences contained various complex syntactic structures. The target sentences were separated from each other by at least one filler item. The first four trials always consisted of filler items to familiarize the participants with the task.

Eight experimental lists were created as follows. The 32 items were randomized such that sets of four consecutive items had one item of each condition (with fillers interspersed). The complementizer was omitted in every other item, counterbalanced across List 1 and 2. Lists 3 and 4 were obtained by reversing the order of presentation in Lists 1 and 2. The randomization procedure was then repeated to generate Lists 5 through 8. Each list was assigned to 16 participants.

**Procedure.** Sentences were presented word by word in a self-paced moving window paradigm (Just, Carpenter, & Woolley, 1982). After each trial, the participants were presented with a Y/N comprehension question to ensure that they were paying attention to the meaning of the sentence. Participants did not receive feedback on their responses. The experiment was conducted online
using a Flash application written by Hal Tily (now at Nuance Communications). Participants took 17 minutes on average to complete the experiment (standard deviation: 4.1 minutes).

**Preprocessing.** Following standard procedure, individual words were excluded if their raw reading times (RT) were less than 100 ms or more than 2000 ms. All RTs were log-transformed to reduce right skew (Baayen & Milin, 2010; Fine & Jaeger, 2014; Frank, 2013). If a word’s log-transformed RT was more than 3 standard deviations higher or lower than the participant's mean log-transformed RT, the word was excluded. RTs were then length-corrected by taking the residuals of a mixed-effects model which had log-transformed RT as the response variable, word length as a fixed effect, and a by-subject intercept and slope (following, e.g., Fine et al., 2013). Again following standard procedure, all trials including fillers were entered into the length-correction model.

Subjects were excluded if their answer accuracy was lower than 75% (two subjects), or if their mean reaction time differed by more than 2.5 standard deviations from the overall mean reaction time across subjects (two subjects). The results reported in what follows are based on the remaining 124 subjects.

**Statistical analysis.** The resulting by-region length-corrected RTs were analyzed using linear mixed-effects models in R (Bates, Maechler, & Bolker, 2012), with crossed random effects for subjects and items. We used a maximal random effect structure: for items, a slope for sentence ambiguity; for subjects, slopes for all of the predictors and their interactions. In case the model fitting procedure did not converge, we removed the random slopes for the highest order interactions and refitted the model. In was occasionally necessary to exclude the random slope for the three-way entropy x surprisal x ambiguity interaction, but generally not necessary to exclude any of slopes for the two-way interactions. P-values for fixed effects were calculated
using model comparison with a simpler model with the same random effect structure but without the fixed effect in question (following Barr, Levy, Scheepers, & Tily, 2013).

**Results**

**Accuracy.** The mean accuracy (including fillers) was 95.8% (standard deviation: 5.6%). To check whether there were accuracy differences between the conditions, a mixed-effects logistic regression model (Jaeger, 2008) was fitted to the responses to the comprehension questions (excluding fillers). There were no significant main effects or interactions (all \( p > 0.1 \), Wald statistic), indicating that accuracy was similar across conditions. For the RT analyses, we analyzed all critical trials, regardless of accuracy.

**Reading times.** Mean RTs averaged within each region are shown in Fig. 4 (see also word-by-word RTs in Fig. S1 in the Supplementary Materials). Following previous work (Garnsey et al., 1997), we split the sentences into five regions: subject, verb, ambiguous region, disambiguating region and “rest” (see (2) above). Length-corrected RTs were averaged for each region, and linear mixed effects models were fitted within each region.\(^5\) The results for all regions are summarized in Table 2.

**Subject region.** No effects reached significance.

**Verb region.** No effects reached significance, including the effect of subcategorization entropy (\( p = 0.43 \)), contrary to the predictions of the two uncertainty-based hypotheses. (Additional follow-up analyses that included a one-word spill-over region in the verb analysis likewise failed to find an effect of entropy.)
Ambiguous region. The subject of the embedded clause (the “ambiguous region”) was read faster in unambiguous sentences than in ambiguous sentences ($p < 0.001$). There was a marginal main effect of SC surprisal ($p = 0.09$). Simple effect analyses showed that RTs were significantly higher for high SC surprisal verbs than for low SC surprisal verbs in unambiguous sentences ($p = 0.03$) but not in ambiguous ones ($p = 0.45$). The interaction was not significant, however ($p = 0.23$). The effect of surprisal in unambiguous sentence may reflect spillover from the complementizer in unambiguous sentences, which is unpredictable after high SC surprisal verbs.

Disambiguating region. This region was read faster in unambiguous sentences ($p < 0.001$). There was a main effect of SC surprisal ($p = 0.03$), as well as significant interaction between SC surprisal and ambiguity in this region ($p = 0.05$), such that the simple effect of SC surprisal in unambiguous sentences was not significant ($p = 0.28$), but the simple effect in ambiguous sentences was significant ($p = 0.007$). This is the signature expectation violation effect observed in previous studies (Garnsey et al., 1997; Trueswell et al., 1993).

Rest region. No effects reached significance.

Discussion

The experiment replicated the SC surprisal effect found in previous studies (Garnsey et al., 1997; Trueswell et al., 1993): In ambiguous sentences, disambiguation in favor of an SC was more costly when the surprisal of an SC given the verb was high. Subcategorization entropy did not significantly affect RTs at the verb (or in any other region of the sentence). The absence of an entropy effect tentatively argues for the surprisal hypothesis and against the competition and entropy reduction hypotheses.
One important caveat to this conclusion is that our experiment was based on the verbs’ subcategorization entropy, that is, on readers’ uncertainty about the syntactic category of the verb. As indicated in the introduction, this quantity does not take into account the reader’s full uncertainty about the parse. We now examine the consequences of replacing subcategorization entropy with full entropy about the syntactic structure of the sentence.

**Full entropy analysis**

In order to assess the effect of the full uncertainty that a comprehender might experience during incremental sentence understanding, we derived full entropy estimates from a probabilistic context free grammar (PCFG) based on the Penn Treebank. These full entropy estimates take into account not only the uncertainty about the next syntactic node, but also the uncertainty about the internal structure of that node (cf. Fig. 3). The procedure is described in detail in Appendix B. We use these full entropy estimates to test the competition and entropy reduction hypotheses, while simultaneously controlling for surprisal (estimated from the same PCFG).

Fig. 5 summarizes the full entropy, entropy reduction and surprisal estimates derived from the PCFG, averaged within each of the four conditions of the factorial design (high vs. low SC surprisal x high vs. low subcategorization entropy). We go through each region separately. For each region, we begin by describing the relation between the factorial design and the PCFG-derived estimates. We then describe the predictions of the competition and entropy reduction hypotheses for RTs (see Table 3 for an overview). Finally, we analyze the effects of entropy and entropy reduction on RTs. We assess the effects of each of the variables in a separate model. Instead of the factorial surprisal and subcategorization entropy predictors, the models included continuous PCFG-based surprisal and one of the full entropy measures, as well as an interaction term. When not mentioned otherwise, all analyses contained the full random effect structure.
INSERT TABLE 3 HERE.

**Subject region**

Since the sentences did not differ in entropy at the subject region, neither the competition nor the entropy reduction hypothesis made any predictions for this region.

**Verb region**

Full entropy in this region is somewhat higher for verbs with high subcategorization entropy (Fig. 5). However, this difference is dwarfed by the substantial correlation between SC surprisal and full entropy: Verbs that are more likely to be followed by an SC (i.e., verbs with lower SC surprisal) have higher full entropy. This correlation, which may be unexpected at first blush, stems from the fact that full entropy at a given point in the derivation is calculated as the sum of single-step entropy and the expected full entropy of the structures that can be derived at that point (see Appendix B for details). SCs have many more potential internal structures than NPs or preposition phrases, and therefore higher internal entropy; when the probability of an SC is high, full entropy is dominated by the internal entropy of an SC (Fig. 6).

Entropy and entropy reduction are perfectly inversely correlated in this region, since entropy before the verb is identical across items. The competition hypothesis predicts RTs to be positively correlated with entropy. The entropy reduction hypothesis predicts the opposite (i.e., a negative correlation with entropy).

INSERT FIGURE 5 HERE.

INSERT FIGURE 6 HERE.

**Results:** Unlike in the subcategorization entropy analysis, where entropy was not a significant predictor of RTs, higher full entropy at the verb predicted faster RTs ($p = 0.05$). This
effect is in the direction predicted by the entropy reduction hypothesis. No other effects reached significance.

**Ambiguous region**

The high internal entropy of SCs continues to play a major role in the full entropy profile of the following regions as well. In unambiguous sentences, the sequence of *that* + determiner increases the probability of an SC to essentially 1 (modulo uncertainty due to grammatical errors or noise), causing full entropy to rise sharply. In ambiguous sentences, on the other hand, some of the probability mass is reserved to the relatively low-entropy NP complement, resulting in overall lower entropy. Counterintuitively, then, the competition hypothesis predicts higher RTs in unambiguous sentences than in ambiguous sentences in this region. The entropy reduction hypothesis, on the other hand, predicts higher RTs in ambiguous than unambiguous sentences: Entropy decreases in ambiguous sentences, leading to some processing cost, whereas it increases in unambiguous sentences (recall that an increase in entropy is not predicted to affect processing time).

Within ambiguous sentences, the predictions of the competition hypothesis for the ambiguous region are qualitatively similar to its predictions at the verb: The more likely the verb is to be followed by an SC, the higher the full entropy and hence the higher the RTs predicted by the competition hypothesis. The entropy reduction hypothesis predicts that this region should show the mirror image of the effect predicted at the verb: Verbs that strongly predict an SC cause a milder entropy reduction at the verb, but a steeper entropy reduction at the word *the*, which increases the probability of an NP and partly counteracts the verb-specific SC bias. At the ambiguous region of ambiguous sentences, then, both competition and entropy reduction predict the same qualitative effect across items: Lower SC surprisal should lead to increased RTs.
Within unambiguous sentences, neither theory predicts differences in RTs across items: Since the sentence has already been disambiguated in favor of the SC parse, entropy estimates for both this region and the word preceding it (that) are identical across items.

**Results:** Although the correlation was not perfect as in the verb region, entropy and entropy reduction were still highly correlated in the ambiguous region ($r = .93$). Both hypotheses predict Ambiguity to be correlated with entropy and entropy reduction; this was indeed the case (entropy: $r = 0.99$; entropy reduction: $r = -0.89$). As such, we first conducted the analyses collapsing across ambiguous and unambiguous sentences. Entropy reduction correlated with increased reading times ($\beta = 0.041, p < 0.001$), and entropy correlated with decreased reading times ($\beta = -0.043, p < 0.001$). No other effects reached significance. These results are in line with the entropy reduction hypothesis, and opposite to the predictions of the competition hypothesis.

To test whether the effect of the entropy measures was solely carried by Ambiguity, we then focused on ambiguous sentences only (there was no variability across items in unambiguous sentences; see Fig. 7). We again fitted two models, one with entropy and one with entropy reduction (and the interaction of either measure with surprisal). None of the predictors in the entropy reduction model had a significant effect on RTs (surprisal: $p > 0.6$, entropy reduction: $p > 0.4$; interaction: $p > 0.4$). In the entropy model, the main effects were not significant (surprisal: $p > 0.4$; entropy: $p > 0.9$); there was an unexpected interaction between the two variables ($p < 0.03$), which upon closer inspection was revealed to be spurious. These results suggest that the effect of entropy and entropy reduction in this region was only due to the difference between ambiguous and unambiguous sentences.

INSERT FIGURE 7 HERE.
**Disambiguating region**

When readers reach the disambiguating region, they have ample evidence that the high-entropy SC parse is the correct one. This leads to high overall entropy, regardless of whether the sentence was originally ambiguous or unambiguous and of expectations derived from the verb’s subcategorization bias. Consequently, the competition hypothesis predicts no difference across sentence types at this region.

Conversely, the entropy reduction hypothesis predicts that the disambiguating region should be read *faster* in ambiguous compared to unambiguous sentences. This somewhat counterintuitive prediction deserves some elaboration. When ambiguous sentences are disambiguated in favor of the SC parse, entropy increases sharply because of the higher internal entropy of SCs. The degree to which it increases depends on the verb’s subcategorization bias; however, the entropy reduction hypothesis predicts processing difficulty only for *decreases* in entropy, and hence never predicts any processing cost at the disambiguation region of ambiguous sentences. For unambiguous sentences, on the other hand, readers already know that the complement is an SC. This means that entropy will go down at the first word of the disambiguating region, because on average additional words reduce entropy (see also Fig. 5). This reduction in entropy, however mild, entails *some* processing cost, compared to no processing cost at all in ambiguous sentences.

Neither hypothesis predicts a difference across items in this region, even within ambiguous sentences. As discussed above, while entropy before the disambiguation point does differ across items, it always *increases* at the first word of the disambiguating region; the entropy reduction hypothesis therefore does not predict any RT difference across items at this word. The differences in entropy before the disambiguation point do not affect the predictions of the
competition hypothesis either, since this hypothesis does not take history into account. From the
disambiguation point on, all items have identical parses, and therefore identical word-by-word
entropy and entropy reduction estimates. Consequently, neither hypothesis predicts any
difference at the second and third word of the disambiguating region (recall that we are assuming
noise free processing, for simplicity’s sake).

Results: Following our analyses for the ambiguous region, we excluded the Ambiguity
factor from the analysis (i.e., collapsed all sentences). The entropy reduction model did not
converge when the interaction between entropy reduction and surprisal was included; we thus
repeated that analysis with only main effects for these two predictors. In both models, surprisal
had a significant effect ($\beta = 0.05, p = 0.003$ with entropy; $\beta = 0.07, p = 0.008$ with entropy
reduction), but the effects of entropy and entropy reduction were not significant, and neither was
the interaction between entropy and surprisal ($ps > 0.2$). The items did not vary sufficiently in
entropy or entropy reduction in this region to warrant follow-up analysis within each sentence
type (Fig. 7).

Rest region

Both surprisal and entropy were identical across items, in line with the RTs on this region
(which therefore do not distinguish between the competing entropy hypotheses).

Summary of full entropy analyses

The PCFG-based analyses reliably replicated the surprisal effect in the disambiguating
region that was already observed in the factorial analysis, suggesting that our probability
estimates were adequate. We found a significant effect of entropy following the verb on RTs at
the verb: Higher post-verb entropy correlated with shorter RTs. This is in line with the prediction
of the entropy reduction hypothesis (under the assumption that the subject region was indeed
uncertain (uninformative), but the opposite of the pattern predicted by the competition hypothesis. In the following regions of the sentence, the entropy reduction hypothesis predicted higher RTs at the ambiguous region of ambiguous sentences, but lower RTs at the disambiguating region. RTs at the ambiguous region support the prediction of the entropy reduction hypothesis; RTs at the disambiguation region, on the other hand, are inconsistent with this hypothesis (see also Table 3 and Fig. 5).

**General discussion**

Building on recent evidence that readers maintain expectations over upcoming syntactic structure, this study investigated how readers’ parsing performance is affected by the probability distribution of those expectations, focusing specifically on readers’ uncertainty about upcoming structure. We outlined three hypotheses about the potential role of uncertainty in parsing: under the competition hypothesis, higher uncertainty should result in the activation of multiple structures that compete with each other, thereby slowing down processing (Elman et al., 2005; McRae et al., 1998); under the entropy reduction hypothesis, processing is slowed down by any word that reduces uncertainty (Hale, 2006); and under the surprisal hypothesis, uncertainty does not affect processing cost at all (Hale, 2001; Levy, 2008).

We assessed uncertainty about the parse in two ways: single-step entropy, which quantifies uncertainty about the next derivation step, in this case the category of the verb’s complement (subcategorization frame); and full entropy, which quantifies uncertainty about the syntactic structure of the whole sentence. Single-step entropy did not affect RTs at the verb, but full entropy did: RTs were longer when post-verb full entropy was lower. The direction of the effect is not compatible with (our implementation of) the competition hypothesis, according to which higher entropy should lead to increased competition and slower processing. It is, however,
consistent with the entropy reduction hypothesis: Entropy before the verb was identical across verbs and always higher than entropy after it; if post-verb (full) entropy is high, then, the verb did not reduce entropy by much, and thus is (correctly) predicted to be relatively easy to process. The entropy reduction hypothesis also correctly predicts that the ambiguous region of unambiguous sentences should be read faster, compared to ambiguous sentences. The only prediction of the entropy reduction hypothesis that was explicitly rejected applies to the disambiguation region. Here, the entropy reduction hypothesis predicts that ambiguous sentences should be read faster than unambiguous sentences. The opposite was observed.

The subcategorization frequency of a sentential complement played a surprisingly central role in determining the full entropy profile of the sentence. Full entropy increases whenever the sentence is disambiguated in favor of an SC parse, and post-disambiguation entropy is the same regardless of the probability of an SC parse given the verb. Both competition and entropy reduction therefore fail to predict the increased reading time on the disambiguating region with high surprisal SCs, which was the most robust effect in the reading time study. This indicates that uncertainty-based measures can only supplement surprisal rather than supersede it.

In summary, it seems that (at least) both the surprisal and the entropy reduction hypothesis are required to account for our results, whereas no support for the competition hypothesis was observed.

What is the amount of syntactic lookahead that needs to be taken into account when estimating uncertainty? The current study has evaluated the two ends of the spectrum: $n = 1$ (single-step entropy) and $n = \infty$ (full entropy). Other values of $n$ are also possible (indeed, likely); our finding that full, but not single-step, entropy predicts RT further supports the conclusion that human parsing during reading involves lookahead of at least several derivation
steps. This conclusion is in line with the conclusions of Frank (2013), who experimented with lookahead distances of 1 to 4 steps and found that increasing lookahead increased the extent to which entropy reduction predicted RTs.\textsuperscript{7}

Changing the lookahead distance may qualitatively change the predictions made by the competition and entropy reduction hypotheses. Consider, for example, the predictions of the competition hypothesis (as implemented here) for the ambiguous region. At first blush, one might expect that ambiguity will lead to more competition because of the uncertainty about the category of the complement (cf. the discussion in Green and Mitchell, 2006; Levy, 2008, p. 1152). However, as we have outlined above, there are actually two components that combine to determine uncertainty (competition) at this point in the sentence: the uncertainty about the category of the complement (e.g., whether it is an SC) and the uncertainty about the internal structure of the complement. Under the infinite lookahead assumption, the latter turns out to dominate the former. The competition hypothesis therefore predicts that the disambiguating region will be processed more \textit{slowly} in unambiguous sentences than in ambiguous ones – contrary to our findings (which replicate Kennison, 2001; Pickering & Traxler, 1998).\textsuperscript{8}

However, since much of the large entropy associated with SCs comes from their internal structure, shorter lookahead distances would decrease the relative contribution of the internal structure of SCs to the overall uncertainty experienced in the ambiguous region, bringing the predictions of the competition hypothesis in line with the empirical findings. Determining the appropriate lookahead distance therefore constitutes an interesting question for future computational studies.

It is also important to point out that entropy estimates and the definition of what constitutes a single derivation step may depend on the strategy employed by the parser and on
the precise representation of the grammar. For example, a parser may choose to defer the prediction of an NP or SC category until there is some information supporting either of these categories (in a top-down parser, this strategy could be implemented by applying a right-binarization transform, which underspecifies the category of the complement; Roark & Johnson, 1999). Such a parsing strategy may predict no uncertainty at all at the verb. Furthermore, the grammar representation we employed was based on the Penn Treebank (Marcus et al., 1993), with minimal modifications (see Appendix B). The Penn Treebank has a small nonterminal set (around 20 nonterminals). Recent work suggests that larger nonterminal sets, created by splitting existing symbols into finer-grained categories (e.g., by annotating a node in the tree with the tags of its siblings and parents) provide a more realistic probabilistic model of natural language syntax (Johnson, 1998; Klein & Manning, 2003; Petrov & Klein, 2007; Roark, 2001). Entropy estimates and lookahead distance based on narrower categories are likely to differ significantly from those based on broad categories: Single-step prediction of a narrow category can approximate several stages of prediction of broader categories. More generally, grammatical formalisms that allow for some degree of context-sensitivity (e.g., Kallmeyer, 2010) have been argued to be more adequate models of human language syntax. In future work, it is worth exploring how these grammatical representational assumptions affect entropy estimates in general and the distinction between single-step and full entropy in particular.

**Conclusion**

This study used syntactic expectations induced by individual lexical items to examine the role of uncertainty over expectations in parsing. The results lend some support to the entropy reduction hypothesis (Hale, 2006): Verbs that reduced uncertainty to a greater extent were read more slowly. However, the entropy reduction hypothesis failed to predict the most robust RT
differences between the conditions, differences that were straightforwardly predicted by the surprisal hypothesis (Hale, 2001; Levy, 2008). Our results suggest that predictability (surprisal) and uncertainty both play a role in explaining processing difficulty in sentence processing. The extent to which uncertainty predicted processing difficulty depended on the depth of syntactic lookahead that readers were assumed to perform: Uncertainty was not a significant predictor of RTs when only the syntactic category of the verb’s complement was considered, and became significant only when the internal complexity of the complement was taken into account.

Appendix A: List of materials

Low subcategorization entropy, low SC surprisal

The men discovered (that) the island had been invaded by the enemy.

The women revealed (that) the secret had been exposed by the officials.

The man noticed (that) the mistake had not happened due to negligence.

The woman assumed (that) the blame might have belonged to the driver.

They all indicated (that) the problem might not bother the entire team.

Two people found (that) the equipment should be reported stolen right away.

Some people sensed (that) the conflict should be resolved quickly and peacefully.

Many people guaranteed (that) the loan would be paid off on time.

Low subcategorization entropy, high SC surprisal

The woman determined (that) the estimate had been inflated by the accountant.

Two people heard (that) the album had been criticized in the magazine.

Some people understood (that) the message had not meant much to foreigners.

They all read (that) the newspaper might be going out of business.

The women worried (that) the parents might have become quite restless recently.
Many people advocated (that) the truth should be made public without delay.
The man taught (that) the children should be sheltered from all harm.
The men projected (that) the film would not gross enough in cinemas.

**High subcategorization entropy, low SC surprisal**

They all claimed (that) the luggage had been stolen from the hotel.
Some people regretted (that) the decision had been reached without any discussion.
The men remembered (that) the appointment had not changed since last week.
The women warned (that) the drivers might have drunk too much vodka.
Many people feared (that) the future might not hold hope for them.
The man proposed (that) the idea should be abandoned for financial reasons.
Two people suggested (that) the scene should be filmed right before sunset.
The woman announced (that) the wedding would be postponed until late August.

**High subcategorization entropy, high SC surprisal**

The men forgot (that) the details had been worked out in advance.
The man observed (that) the patient had been sent home too early.
The woman recalled (that) the speech had not gone over very well.
The women answered (that) the questions might be discussed during the meeting.
Some people added (that) the numbers might have decreased since last year.
Two people wrote (that) the interview should be conducted over the phone.
Many people advised (that) the president should be considering further budget cuts.
The men begged (that) the judge would not treat the defendant harshly.

**Appendix B: Grammar definition and estimation**

**Definitions**
A probabilistic context free grammar (PCFG) consists of a set of non-terminals $V$, which includes intermediate categories such as VP (verb phrase) and N (noun); a set of terminal symbols $T$, which represent specific words (e.g., dog); a special start symbol $S$; a set of rule productions of the form $X \rightarrow \alpha$, where $X$ is a nonterminal and $\alpha$ is a sequence of terminals or nonterminals (e.g., $VP \rightarrow V \; NP$); and a function $\rho$ that assigns a probability to each production rule. We define $R(X)$ to be the set of rules rewriting nonterminal $X$.

A PCFG is considered lexicalized if some of its nonterminals include lexical annotations that allow the identity of a lexical head to affect the probability of modifiers that co-occur with it. For example, a lexicalized grammar can have both $\rho(VP_{\text{break}} \rightarrow V_{\text{break}} \; NP) = 0.3$ and $\rho(VP_{\text{hit}} \rightarrow V_{\text{hit}} \; NP) = 0.7$.

We first define the entropy of the next derivation step (single-step entropy). If $a_i \in V$ is a nonterminal (e.g., $VP$), the single-step entropy $h(a_i)$ corresponding to $a_i$ is given by:

$$h(a_i) = - \sum_{r \in R(a_i)} \rho(r) \log_2 \rho(r)$$

Full entropy is defined recursively, as the sum of two terms: the single-step entropy on the one hand and the sum of the expected full entropy of the nonterminals that can be derived from it on the other hand. Formally, the full entropy $H(a_i)$ of nonterminal $a_i$ is given by:

$$H(a_i) = h(a_i) + \sum_{r \in R(a_i)} \rho(r) \sum_{j=1}^{k_r} H(a_{r,j})$$

Where $a_{r,1} \ldots a_{r,k_r}$ are the nonterminals on the right hand side of rule $r$. The closed form formula for the recursion is:

$$H = (I - A)^{-1} h$$
Where \( H = (H_1, \ldots, H_V) \) is the vector of all full entropy values, \( h = (h_1, \ldots, h_V) \) is the vector of all single-step entropy values, \( I \) is a \(|V| \times |V|\) identity matrix, and \( A \) is a \(|V| \times |V|\) matrix, in which the element in row \( i \) and column \( j \) indicates the expected count of instances of nonterminal \( a_j \) resulting from rewriting nonterminal \( a_i \).

**Full entropy estimation**

Word-by-word entropy estimates for our materials were derived using the Cornell Conditional Probability Calculator (Hale, 2013) from a probabilistic context-free grammar estimated from the Penn Treebank. Following standard practice, we removed grammatical role and filler-gap annotations (e.g., NP-SUBJ-2 was replaced by NP). We reduced the size the grammar by removing rules that included punctuation, rules that occurred less than 100 times (out of the total 1320490 nonterminal productions) and rules that had a probability of less than 0.01. These steps resulted in the removal of 13%, 14% and 10% rule production tokens, respectively.

The grammar was unlexicalized, except for verb-specific production rules that captured the differences in subcategorization probabilities among the 32 verbs in the experiment (again based on the Gahl et al., 2004, norms). Half of the probability mass from all (unlexicalized) rules deriving VP was divided among the lexicalized rules. The conditional probability of each rule was proportional to the verb’s frequency. For example, the probability of the rule

\[
VP_{\text{discover}} \rightarrow V_{\text{discover}} \, NP
\]

was defined to be:

\[
\rho(VP_{\text{discover}} \rightarrow V_{\text{discover}} \, NP) = \frac{1}{2} \frac{\text{freq}(\text{discover}) P(NP|\text{discover})}{\sum_i \text{freq}(v_i)}
\]
Unlexicalized grammars read off a treebank have been shown to make excessively strong assumptions of context-freeness, which affects the accuracy of probability estimates derived from such grammars (Johnson, 1998). The adequacy of the grammar is typically improved either by lexicalizing the grammar or by adding contextual information to some of the tags; e.g., the NP tag may be split into NP^VP for an NP whose parent is an VP and NP^S for an NP whose parent is an S (Johnson, 1998; Klein & Manning, 2003). At the same time, the number of nonterminals in the grammar had to be kept to a minimum to enable the use of the closed form full entropy formula, which requires inverting a matrix that has as many rows as the number of nonterminals in the grammar. We therefore only split the tags that were most relevant to the probability estimates derived for the experimental materials. First, the word that is tagged in the Penn Treebank as a preposition (IN) when it occurs as a subordinating conjunction (as in the men discovered that...). This resulted in SCs being erroneously parsed as prepositional phrases. We therefore replaced the generic tag IN with IN[that] in those cases. Second, the grammar assigned implausibly high probabilities to reduced relative readings of the materials (where discovered that the island is attached as a modifier of the men, by analogy with the men discovered by the police). Since the verb in a reduced relative must be a past participle (VBN), we split the VP category into subcategories based on the VP’s leftmost child, e.g., VP_VBD is a VP headed by a past-tense verb (VBD), such that only a VP_VBN can modify a noun. We likewise split SBAR into SBAR[overt] when the SBAR had an overt complementizer and SBAR[none] when it did not.
References


Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.


Association for Computational Linguistics.


Footnotes

1 Another way in which Frank’s (2013) measure of entropy reduction differs from the measure evaluated in the current paper is that he assumes that increased entropy facilitates processing, whereas we follow the original proposal by Hale (2006) in assuming that it does not.

2 Only frames that were compatible with active uses of the verb were considered. Additionally, Gahl and colleagues distinguish frames that include participles (Lola looked up from her book for the intransitive frame) from frames that do not (we watched attentively); we ignored this distinction for the purposes of calculating subcategorization frame entropy.

3 This meant that sentence subjects were repeated across items (4 times each). This choice was made in order to avoid more informative subjects. It is, however, theoretically possible that participants implicitly learned over the course of the experiment that these eight subjects were always predictive of an SC. We report follow-up analyses that rule out this possibility in the Supplementary Materials.

4 Most of the verbs were ambiguous between past tense and passive participle interpretations. In principle, this allows a reduced relative continuation; however, this continuation is very rare (<1%, cf. Fine et al., 2013), and implausible for most of the verbs (e.g., the men discovered the island had been invaded cannot be interpreted as having the same structure as the men sent the letter were arrested).

5 We repeated the analyses with SC surprisal and subcategorization entropy as continuous predictors. The qualitative pattern of results was similar, with one exception: In the ambiguous region, the simple effect of SC surprisal for unambiguous sentences, while in the expected direction, was not significant ($p = 0.14$).
This interaction was only significant because of a single item, the verb *worry*, for which the surprisal of the noun phrase in the ambiguous region was more than 3 standard deviation higher than the average surprisal across items. This is due to the high probability of prepositional phrase arguments for the verb *worry*, e.g. *worry about their health*. When the item containing the verb *worry* is removed from the analysis, the interaction is no longer significant. None of the other results were affected by the removal of this item.

We note that the lookahead distance in Frank (2013) is not directly comparable to ours. Frank calculated entropy based on word predictions derived from a connectionist network; a lookahead of $n = 4$ in that model corresponds to predicting the next four words. Conversely, our model predicts PCFG rewrite rules, not words; multiple PCFG derivation steps may be required to predict each word (and vice versa), such that four words can correspond to three, five or eight PCFG rewrite rules.

Alternatively, the longer reading times in the ambiguous region of ambiguous sentences may reflect a spillover effect, resulting from the fact that the ambiguous region follows a frequent function word in unambiguous sentences (*that*), but an infrequent content word in ambiguous sentences (namely, the verb; cf. Clifton & Staub, 2008).
Table 1

*Lexical variables*

<table>
<thead>
<tr>
<th>Condition</th>
<th>Entropy</th>
<th>SC-surprisal</th>
<th>Frequency (log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low entropy / Low surprisal</td>
<td>1.13 (0.08)</td>
<td>1.52 (0.55)</td>
<td>3.64 (1.56)</td>
</tr>
<tr>
<td>Low entropy / High surprisal</td>
<td>1.09 (0.18)</td>
<td>4.15 (1.03)</td>
<td>4.31 (2.01)</td>
</tr>
<tr>
<td>High entropy / Low surprisal</td>
<td>1.7 (0.12)</td>
<td>1.58 (0.45)</td>
<td>3.6 (1.24)</td>
</tr>
<tr>
<td>High entropy / High surprisal</td>
<td>1.68 (0.17)</td>
<td>3.86 (0.79)</td>
<td>3.85 (1.25)</td>
</tr>
</tbody>
</table>

*Note.* Mean subcategorization entropy, SC surprisal and log-transformed frequency in each of the conditions of the factorial design. Standard errors are shown in parentheses.
Table 2

*Statistical analysis of factorial design*

<table>
<thead>
<tr>
<th></th>
<th>Subject</th>
<th>Verb</th>
<th>Ambiguous</th>
<th>Disambiguating</th>
<th>Rest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>0.67</td>
<td>0.86</td>
<td>-0.79</td>
<td>0.98</td>
<td>0.45</td>
</tr>
<tr>
<td>Surprisal</td>
<td>0.82</td>
<td>1.66</td>
<td>1.6 (.)</td>
<td>2.05 (*)</td>
<td>0.96</td>
</tr>
<tr>
<td>Ambiguity</td>
<td>0</td>
<td>-0.49</td>
<td>5.95 (***)</td>
<td>3.96 (***</td>
<td>-0.55</td>
</tr>
<tr>
<td>Entropy x Surprisal</td>
<td>-0.12</td>
<td>-1.08</td>
<td>-1.78</td>
<td>-0.25</td>
<td>1.08</td>
</tr>
<tr>
<td>Entropy x Ambiguity</td>
<td>-1.6</td>
<td>-0.96</td>
<td>-0.96</td>
<td>1.19</td>
<td>1.28</td>
</tr>
<tr>
<td>Surprisal x Ambiguity</td>
<td>-0.89</td>
<td>-0.01</td>
<td>-1.13</td>
<td>1.89 (.)</td>
<td>1.79 (.)</td>
</tr>
<tr>
<td>Entropy x Ambiguity x Surprisal</td>
<td>-0.13</td>
<td>-0.87</td>
<td>-0.35</td>
<td>-0.13</td>
<td>1.19</td>
</tr>
</tbody>
</table>

*Note.* The table shows t statistics from a linear mixed-effects regression model.
Table 3.

*Predictions made by uncertainty-based theories (based on full entropy)*

<table>
<thead>
<tr>
<th></th>
<th>Verb</th>
<th>Ambiguous</th>
<th>Disambiguating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Competition</strong></td>
<td>Lower SC surprisal,</td>
<td>Longer RTs in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Higher subcat entropy</td>
<td>unambiguous than in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➔ Higher full entropy</td>
<td>ambiguous sentences ✗</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➔ Longer RTs ✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Entropy</strong></td>
<td>Lower SC surprisal,</td>
<td>Shorter RTs in</td>
<td></td>
</tr>
<tr>
<td><strong>Reduction</strong></td>
<td>Higher subcat entropy</td>
<td>unambiguous than in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➔ Higher full entropy</td>
<td>ambiguous sentences ✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>➔ Lower full entropy</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>reduction ➔ Shorter</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RTs ✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Within ambiguous sentences only:*

<table>
<thead>
<tr>
<th></th>
<th>Verb</th>
<th>Ambiguous</th>
<th>Disambiguating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Competition</strong></td>
<td>Lower SC surprisal,</td>
<td>Longer RTs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Higher subcat entropy</td>
<td>➔ Longer RTs ✗</td>
<td></td>
</tr>
<tr>
<td><strong>Entropy</strong></td>
<td>Lower SC surprisal,</td>
<td>Longer RTs</td>
<td></td>
</tr>
<tr>
<td><strong>Reduction</strong></td>
<td>Higher subcat entropy</td>
<td>➔ Longer RTs ✗</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Predictions made by the competition and entropy reduction hypotheses for the three main regions of the materials (the hypotheses do not make any predictions for the subject and “rest” regions). Predictions are shown both for the whole data set, focusing on the comparison between ambiguous and unambiguous sentences, and specifically for ambiguous sentences. Cells are shaded whenever a hypothesis does not predict any RT difference in a region. Predictions confirmed by the results are marked with ✓; predictions not found confirmed are marked with ❋; predictions rejected by the results are marked with ✗.
Figure 1. Incremental parses with single-step prediction while reading the sentence *he accepted the proposal was wrong*. (a) and (b) represent parses after the verb has been read and its complement type has been predicted. (c) and (d) represent the parses after the ambiguous region *the proposal* has been incorporated into (a) and (b) respectively.
Figure 2. Entropy values for four examples of subcategorization distributions: (a) two balanced frames; (b) two unbalanced frames; (c) three balanced frames; (d) three unbalanced frames.

When probability is evenly distributed across subcategorization frames, verbs with more frames have higher entropy (compare (a) to (c)). For the same number of frames, entropy is lower the less balanced the distribution (compare (a) to (b), or (c) to (d)).
Figure 3. Predicting syntactic structure an unlimited number of derivation steps ahead: In addition to predicting the category of the complement (NP or SC) the reader predicts its internal structure. (a) prediction of a simple determiner + noun phrase (he accepted the present); (b) prediction of a noun phrase with an adjective (he accepted the nice present); (c) prediction of an SC with an intransitive verb (he accepted that he lost); (d) prediction of an SC with a transitive verb (he accepted that he lost her).
Figure 4. Mean reading times (RTs). Error bars show bootstrapped 95% confidence intervals.
Figure 5. Word-by-word entropy, entropy reduction and surprisal predictions derived from the probabilistic context free grammar based on the Penn Treebank. Predictions are averaged within each of the four conditions of the factorial design.
Figure 6. Illustration of the effect of the internal entropy of the verb’s complements on full entropy after the verb. The internal entropy of an SC is much higher than both verbs’ subcategorization entropy and the internal entropy of an NP; the most important predictor of full entropy in this case is therefore the probability of an SC (the specific values of internal entropy and subcategorization probabilities in the figure are for illustration purposes only).

\( h + 0.1H_{NP} + 0.9H_{SC} = 46.8 \)  
\( h + 0.5H_{NP} + 0.5H_{SC} = 33 \)
Figure 7. Variability across items of PCFG-derived variables (entropy, entropy reduction and surprisal).
Additional analyses of reading times

Word-by-word results. Visual inspection of the word-by-word means suggests that there are few differences between conditions in unambiguous sentences, and larger differences in ambiguous sentences. The largest visual divergence between conditions was at the second word of the ambiguous region (island in the example), where higher entropy led to shorter reading times; and at the second word of the disambiguating region (been), where higher entropy led to longer reading times, particularly when SC surprisal was high. Linear mixed-effects models show that in the second word of the ambiguous region, the main effect of entropy does not reach significance (\(p = 0.11\)), and neither do the main effect of surprisal (\(p = 0.59\)) and the interaction with surprisal (\(p = 0.1\)) (though there is no clear reason to expect an effect of SC surprisal at this point). In the second word of the disambiguating region, higher entropy led to higher reading times, though again the effect did not quite reach significance (\(p = 0.07\)). There was a robust effect of surprisal (\(p = 0.004\)), and no interaction (\(p = 0.38\)).

Order effects. In the experimental materials, SC surprisal and subcategorization entropy were orthogonally manipulated. Outside of the experimental context, however, these two quantities are not independent. In addition, in naturalistic language high surprisal complements are not as likely as low surprisal ones. In light of recent evidence that participants adapt their syntactic expectations to the statistical properties of the environment (Fine et al., 2013), we investigated whether the effects of surprisal and entropy change over the course of the experiment, focusing on the ambiguous sentences. There were robust main effects of list
position, such that participants became faster overall in later trials; this effect was particularly strong in the disambiguating region (subject: $\beta = -0.004$; verb: $\beta = -0.003$; ambiguous: $\beta = -0.005$; disambiguating: $\beta = -0.01$; rest: $\beta = -0.005$; all $p < 0.001$). Crucially, however, the order effect did not interact with SC surprisal ($t = -0.22$) or entropy ($t = -1.58$). We conclude that there is no evidence that participants adapted their expectations over the course of the experiment.
Figure S1. Word-by-word reading times.