COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–30
Copyright © 2015 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12274

Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions

Tal Linzen,[a] T. Florian Jaeger[b]

[a] *Department of Linguistics, New York University*
[b] *Department of Brain and Cognitive Sciences, Department of Computer Science, and Department of Linguistics, University of Rochester*

Received 9 October 2014; received in revised form 28 April 2015; accepted 4 May 2015

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* about those expectations. In a self-paced reading study, we use lexical subcategorization distributions to factorially manipulate both the strength of expectations and the uncertainty about them. 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 increased reading times (RTs). We additionally replicated previously observed effects of expectation violation (surprisal), orthogonal to the effect of uncertainty. This suggests that both surprisal and uncertainty affect human RTs. We discuss the consequences for theories of sentence comprehension.

Keywords: Sentence processing; Uncertainty; Prediction; Entropy reduction; Surprisal; Competition

1. Introduction

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; Federmeier, 2007; Hale, 2001; Levy, 2008). These

Correspondence should be sent to Tal Linzen, Department of Linguistics, New York University, New York, NY 10003. E-mail: linzen@nyu.edu

predictions are probabilistic: There is a continuous relationship between predictability and processing difficulty (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008; Jennings, Randall, & Tyler, 1997; McDonald & Shillcock, 2003a,b). 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 have 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, b). 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, d). 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

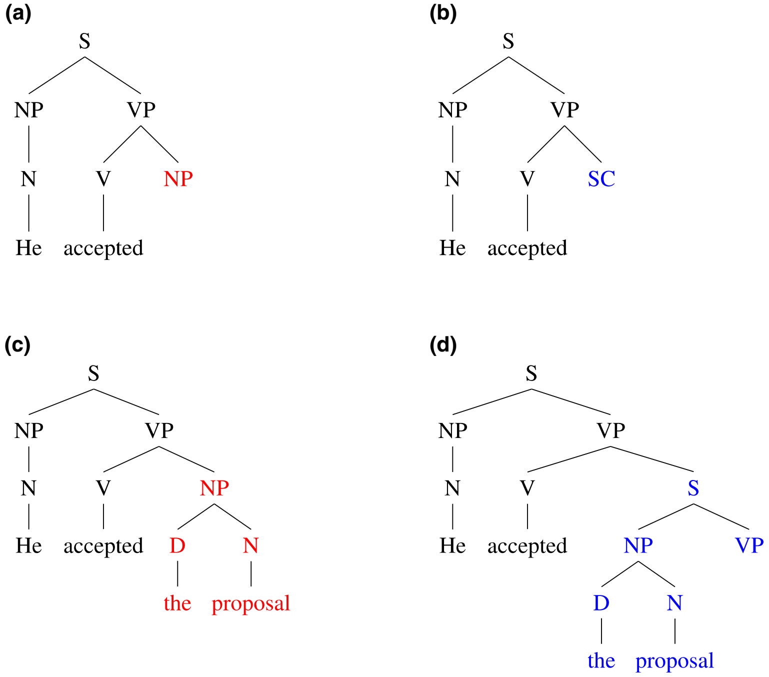

Fig. 1. Incremental parses with single-step prediction while reading the sentence *he accepted the proposal was wrong*. (a) and (b) represent parses after the verb has been read and its complement type has been predicted. (c) and (d) represent the parses after the ambiguous region *the proposal* has been incorporated into (a) and (b), respectively.

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, & Kello, 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 *forget* 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 as follows:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Entropy is higher the more potential parse completions there are and the more uniformly 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 three equiprobable options the entropy is 1.58 bits. If one of the three 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 two 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; Yun, Chen, Hunter, Whitman, & Hale, 2015). Under this hypothesis, an increase in uncertainty does not affect processing; that is, if H_i is the entropy at the *i*-th word, then entropy reduction at the *i*-th word is given by $\max\{H_i - H_{i-1}, 0\}$.

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

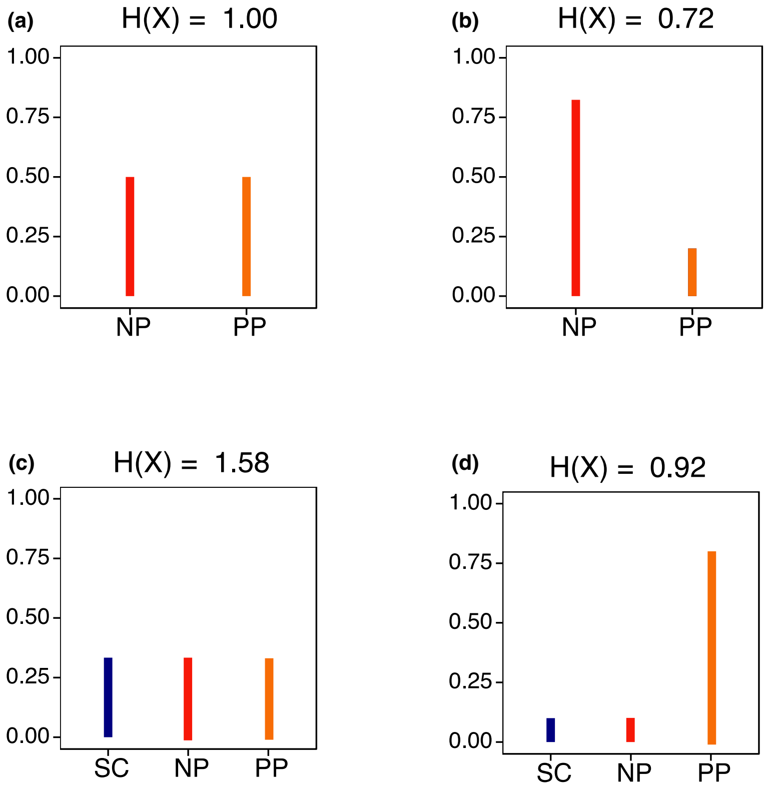

Fig. 2. Entropy values for four examples of subcategorization distributions: (a) two balanced frames; (b) two unbalanced frames; (c) three balanced frames; (d) three unbalanced frames. When probability is evenly distributed across subcategorization frames, verbs with more frames have a higher entropy (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)).

(*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*).

A number of recent studies have begun to explore the effect of uncertainty on reading times (RTs; Frank, 2013; Hale, 2003, 2006; Roark, Bachrach, Cardenas, & Pallier, 2009; Wu, Bachrach, Cardenas, & Schuler, 2010; Yun et al., 2015) and neural measures (Frank, Otten, Galli, & Villioco, 2015; Willems, Frank, Nijhof, Hagoort, & van den Bosch, 2015), some with positive results. However, as we detail next, the conclusions that can be drawn from these studies are limited by the considerable variability between them, as well as lack of critical controls. This state of affairs has led others to question the extent to which these studies provide support for the role of uncertainty-based hypothesis, in

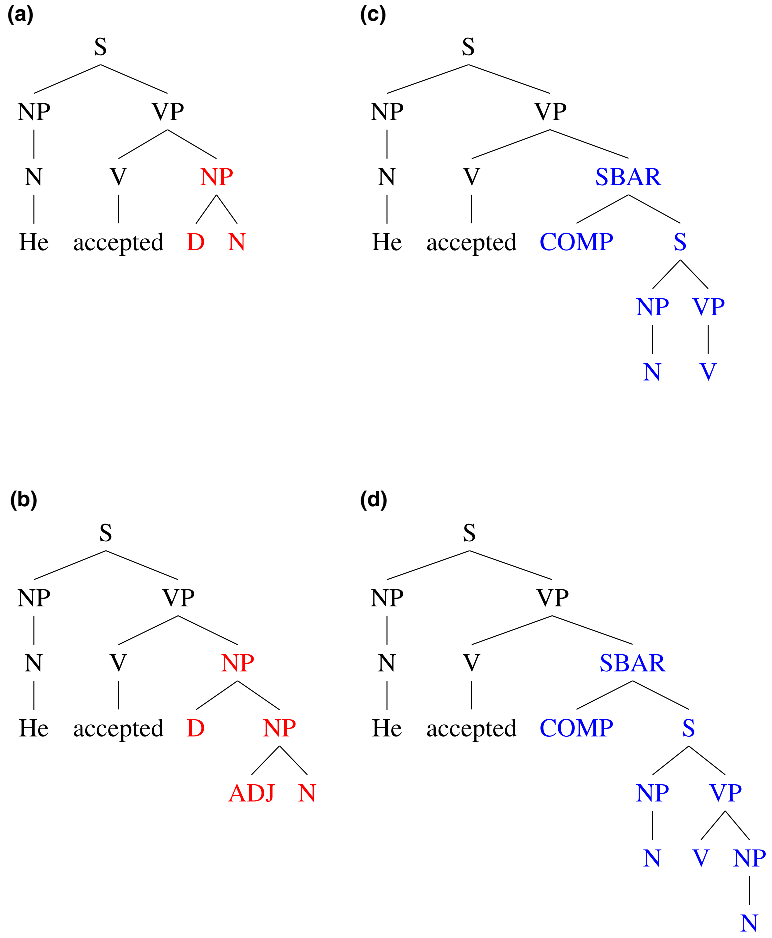

Fig. 3. Predicting syntactic structure an unlimited number of derivation steps ahead: In addition to predicting the category of the complement (NP or SC), the reader predicts its internal structure. (a) Prediction of a simple determiner + noun phrase (*he accepted the present*); (b) prediction of a noun 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*).

particular the entropy reduction hypothesis (Levy & Gibson, 2013). We begin by summarizing the previous studies; we then present the motivation for the current study.

2. Previous work on entropy in sentence processing

One class of studies has demonstrated that the entropy reduction hypothesis can predict qualitative findings from the sentence processing literature, such as the processing difficulty at the disambiguation point in garden path sentences (*the horse raced past the barn* *fell*; Hale, 2003) or the asymmetry between object and subject relative clauses (Yun et al., 2015). These predictions have not been evaluated on empirical RTs and the studies 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 time corpora (Frank, 2013; Roark et al., 2009; Wu et al., 2010). Roark et al. (2009) found a positive effect of entropy over syntactic parses, supporting the competition hypothesis. Support for the entropy reduction hypothesis comes from two studies: Wu et al. (2010) found that entropy reduction had a positive effect on RTs, though only for closed class words, and Frank (2013) found a positive effect of entropy reduction for all words.

Several factors limit the information that can be gained from these studies. Foremost, most previous studies did not directly compare the competition and entropy reduction hypothesis. This is problematic since entropy and entropy reduction effects will tend to be highly correlated (this is confirmed in the present study). A comparison across these previous studies is further made difficult by methodological differences between these studies. First, previous work has used a wide variety of syntactic models, ranging from connectionist networks (Frank, 2013) through unlexicalized (Hale, 2003) and lexicalized probabilistic context-free grammars (Roark et al., 2009) to Hierarchical Hidden Markov Models (Wu et al., 2010). It is thus unclear to what extent the divergent results of previous studies are due to differences in representational assumptions.

Second, the studies differed substantially in the way in which entropy (and thus entropy reduction) was calculated. For example, Hale (2003) calculated entropy over full sentences (full entropy); this measure includes uncertainty both about the analysis of the part of the sentence read so far and about the rest of its structure. This contrasts with the entropy measure employed by Frank (2013), which captures uncertainty about a few upcoming words (that study used a neural network that does not represent syntactic ambiguity and therefore never has any uncertainty about the correct analysis of the past). The approach taken by Roark et al. (2009) falls between these two extremes, in that it captures uncertainty about the analysis of the part of the sentence read so far, as well as uncertainty about the syntactic structure that would need to be constructed to accommodate a single upcoming word.

Third, some authors evaluated the entropy reduction hypothesis while assuming that increases in entropy *facilitate* processing (Frank, 2013), contrary to formulation of the entropy reduction hypothesis (Hale, 2006). In some cases, effects of entropy reduction and of entropy may differ only in their sign (see below), making it impossible to distinguish between the competition and entropy reduction hypothesis as formulated by Hale (2006).

Finally, most previous studies examined processing difficulty across different types of structures, potentially confounding structural differences (and, e.g., correlated differences in memory cost, cf. Gibson, 1998) with differences in entropy. Taken together, these differences limit the conclusions that can be drawn from previous work about the role of uncertainty during language processing.

3. Contribution of the paper

As mentioned above, some researchers have questioned the viability of entropy reduction as a predictor of processing, in part because of the complicating factors listed above (Levy & Gibson, 2013). The goal of the current study is to assess the effects of entropy and entropy reduction in the same materials, within the same syntactic framework, while avoiding structural confounds.

Following Roark et al. (2009), we evaluate the predictions of uncertainty-related hypotheses on human RTs. 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 from entropy *reduction* effects. Positive correlations between entropy and reading times are predicted by the competition hypothesis, whereas positive correlations between entropy reduction and reading times are predicted by the entropy reduction hypothesis (cf. Hale, 2006). Second, we compare uncertainty in single-step prediction and uncertainty in full prediction. The entropy reduction hypothesis (as formulated in Hale, 2006) predicts only that the latter should be correlated with RTs. This distinction is thus critical, but so far it has received little attention in the literature, which has employed a variety of different measures of entropy and entropy reduction under the implicit assumption that they are interchangeable. The results of the current study show that uncertainty in single-step prediction and uncertainty in full prediction can differ dramatically and exhibit qualitatively different correlations with RTs.

To avoid the potential confounds associated with comparing RTs across constructions, we manipulate syntactic uncertainty while keeping constant syntactic structure and associated confounding factors such as memory cost. We do so by varying the syntactic expectations induced by specific lexical items, as in the case of *accept* compared to *forget* above. Much previous work has shown that syntactic expectations in human language processing are sensitive to lexical information (Garnsey et al., 1997; Linzen, Marantz, & Pylkkänen, 2013; McRae et al., 1998; Spivey-Knowlton & Sedivy, 1995; Trueswell et al., 1993).

We first describe an RT 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 predicts RTs.

4. Reading time 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*). We refer to this factor as Ambiguity.

(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*

For consistency with previous studies, we refer to the subject of the embedded clause (*the island*) as the ambiguous region even when the sentence is unambiguous; likewise, we refer to the verbal complex *has been invaded* as the disambiguating region in both sentence types. 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).

To factorially cross subcategorization entropy and surprisal, we leveraged the fact that many verbs occur with frames other than SC and NP.[1] For example, *find* and *propose* have a similar SC subcategorization probability (0.22 and 0.25, respectively) and thus similar SC surprisal. But *propose* occurs with multiple other frames (NP: 0.57; infinitives: 0.14), whereas for *find* the NP frame is the only alternative to SC that occurs with a substantial probability (0.72). As a result, *propose* has higher subcategorization entropy than *find* (1.56 vs. 1.09).

4.1. Predictions

In summary, the RT study used a 2 × 2 × 2 design: Ambiguity × subcategorization entropy × SC surprisal. If increased uncertainty causes a processing slowdown, as argued by the competition hypothesis, RTs 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 RTs on verbs with lower subcategorization entropy: Since entropy before the verb is matched across conditions (see below), verbs with lower subcategorization entropy reduce uncertainty more than verbs with higher subcategorization entropy. Finally, surprisal (Hale, 2001) predicts 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.

4.2. Method

4.2.1. Participants

A total of 128 participants were recruited through Amazon Mechanical Turk and were paid \$1.75 for their participation. Participants took 17 minutes on average to complete the experiment (*SD* = 4.1 minutes).

4.2.2. Materials

We selected 32 verbs, eight in each of the cells of the 2 × 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). Gahl et al. (2004) classify subcategorization frames into six categories: transitive (*Klaus adore cookies*), intransitive (*we watched attentively)*, quote (*he said* "*that's fine by me*"), finite sentential complement (*Trent yelled (that) the road was in sight*), infinitival complement (*she wanted to share her insight with others*), and "other."[2] 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.

In the next step, 32 sentence 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 main 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 main 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).[4] The frequency of this 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.

Table 1
Lexical variables

Condition	Subcategorization Entropy	SC-Surprisal	Frequency (log)
Low entropy/low surprisal	1.13 (0.08)	1.52 (0.55)	3.64 (1.56)
Low entropy/high surprisal	1.09 (0.18)	4.15 (1.03)	4.31 (2.01)
High entropy/low surprisal	1.7 (0.12)	1.58 (0.45)	3.6 (1.24)
High entropy/high surprisal	1.68 (0.17)	3.86 (0.79)	3.85 (1.25)

Notes. Mean subcategorization entropy, SC surprisal, and log-transformed frequency of the main verb in each of the conditions of the factorial design. We use Entropy to refer to high versus low subcategorization frame entropy (i.e., single-step entropy at the verb) and Surprisal to refer to high versus low sentential complement surprisal. Standard errors are shown in parentheses.

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 Lists 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.

4.2.3. 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. Participants took 17 minutes on average to complete the experiment (*SD* = 4.1 minutes).

4.2.4. Preprocessing

Following standard procedure, individual words were excluded if their raw RTs were less than 100 ms or more than 2,000 ms. All RTs were log-transformed to reduce right skew (Baayen & Milin, 2010; Fine & Jaeger, unpublished data; Frank, 2013). If a word's log-transformed RT was more than 3 *SD*s 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.

Two subjects were excluded because their answer accuracy was lower than 75%. The results reported in what follows are based on the remaining 126 subjects.

4.2.5. 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.[5] The *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).[6]

4.3. Results

4.3.1. Accuracy

Comprehension accuracy, including on fillers, was high (M = 95.8%, SD = 5.6%). To test whether accuracy differed between 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 ps > .1, Wald statistic), indicating that accuracy was similarly high across conditions. For the RT analyses, we analyzed all critical trials, regardless of accuracy.

4.3.2. Reading times

Mean RTs averaged within each region are shown in Fig. 4 (see also word-by-word RTs in Fig. 8 in Appendix C). Following previous work (Garnsey et al., 1997), we split the sentences into five regions: subject, verb, ambiguous region, disambiguating region, and the rest of the sentence (see (2) above). Only the first four regions were statistically analyzed. Length-corrected RTs were averaged for each region of a given trial, and linear mixed effects models were fitted within each region. The results for all regions are summarized in Table 2.

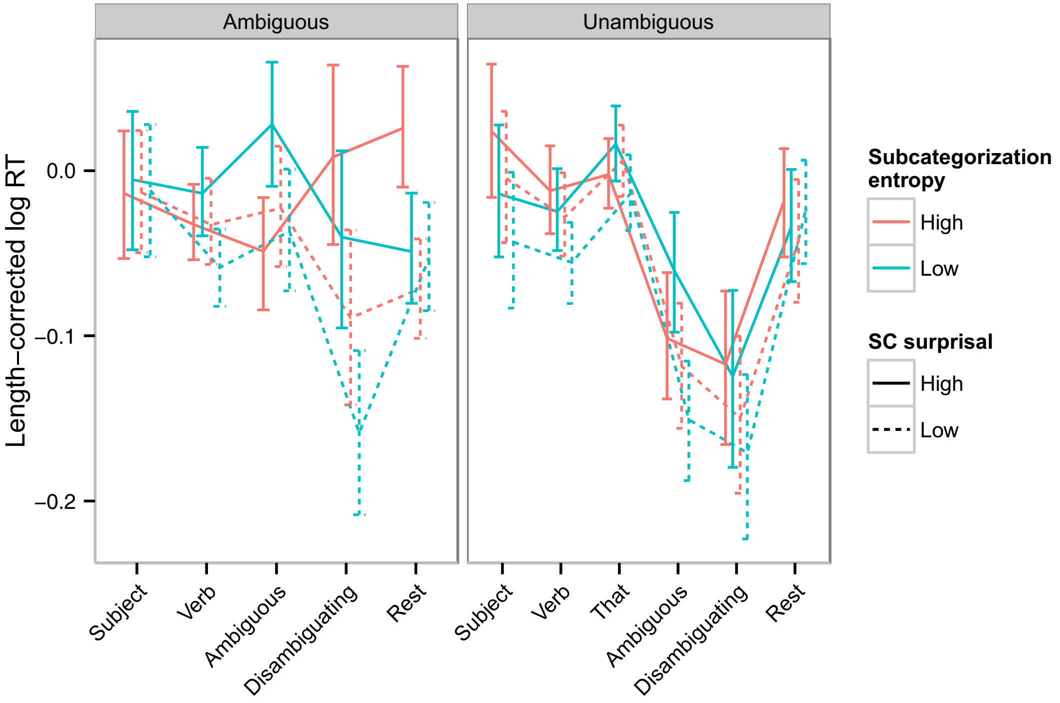

Fig. 4. Mean reading times (RTs). Error bars show bootstrapped 95% confidence intervals.

4.3.2.1. Subject region: No effects reached significance (all ps > .1).

4.3.2.2. Verb region: Subcategorization entropy did not have a significant effect on RTs ($p > .5$). RTs in this region therefore do not support either the competition hypothesis or the entropy reduction hypothesis in their single-step form. Additional follow-up analyses that included a one-word spillover region in the verb analysis likewise failed to find an effect of entropy. Unexpectedly, low SC surprisal verbs were read more slowly, though this difference was only marginally significant ($p = .09$). None of the theories considered here predict an effect of SC surprisal at the verb: At the verb, readers do not yet know the category of the upcoming syntactic complement. We return to this trend when we discuss the full entropy analysis below, as it offers an alternative explanation for this apparent effect of SC surprisal.

4.3.2.3. Ambiguous region: The subject of the embedded clause (the “ambiguous region”) was read faster in unambiguous sentences than in ambiguous sentences ($p < .001$). There was a marginal main effect of SC surprisal ($p = .1$). Simple effect analyses showed that RTs were significantly higher for high SC surprisal verbs than for low SC surprisal verbs in unambiguous sentences ($p = .04$) but not in ambiguous ones ($p > .4$). The interaction was not significant, however ($p > .2$). The effect of surprisal in unambiguous sentences may reflect spillover from the complementizer in unambiguous sentences, which is unpredictable after high SC surprisal verbs.

The main effect of subcategorization entropy was not significant ($p > .3$). There was a significant interaction between entropy and surprisal ($p = .04$). Simple effect analyses, collapsing across the two levels of Ambiguity, showed that high SC surprisal verbs were associated with higher RTs when their subcategorization entropy was low ($p = .02$) but not when it was high ($p > .7$). The three-way interaction between surprisal, entropy, and Ambiguity did not approach significance ($p > .5$).

Table 2
Statistical analysis of factorial design

	Subject	Verb	Ambiguous	Disambiguating
Entropy	0.6	0.81	−0.95	1.03
Surprisal	0.87	1.61†	1.58†	2.17*
Ambiguity	−0.02	−0.55	5.71***	3.84***
Entropy × Surprisal	−0.07	−1.22	−1.94*	−0.32
Entropy × Ambiguity	−1.73†	−1.07	−1.00	1.00
Surprisal × Ambiguity	−1.01	−0.02	−1.11	1.87†
Entropy × Ambiguity × Surprisal	0.2	−0.8	−0.18	−0.14

Notes. The table shows t statistics from a linear mixed-effects regression model. Entropy refers to high versus low subcategorization frame entropy (i.e., single-step entropy at the verb); Surprisal refers to high versus low sentential complement surprisal; and Ambiguity is positive for ambiguous sentences and negative for unambiguous ones. Legend: ***$p < .001$, *$p < .05$, †$p < .1$.

4.3.2.4. Disambiguating region: This region was read faster in unambiguous sentences ($p < .001$). There was a main effect of SC surprisal ($p = .03$), as well as a marginally significant interaction between SC surprisal and ambiguity in this region ($p = .06$), such that the simple effect of SC surprisal in unambiguous sentences was not significant ($p > .2$), but the simple effect in ambiguous sentences was ($p = .007$). This is the signature expectation violation effect observed in previous studies (Garnsey et al., 1997; Trueswell et al., 1993). The main effect of entropy did not reach significance, and neither did any of the interactions between entropy and other predictors (*p*s > .3).

4.4. 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 does not support either the competition or the 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's complement. 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.

5. 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. With the exception of the main verbs, whose lexically specific subcategorization probabilities are of direct relevance to this study, the grammar was unlexicalized: Its rules only made reference to parts of speech (syntactic categories) rather than individual lexical items. In contrast with the substantial transformations applied to the grammar in some state-of-the-art parsers (Johnson, 1998; Petrov & Klein, 2007), the grammar we used was very close to an untransformed "vanilla" PCFG. We made this decision to keep the grammar reasonably small, since computation of full entropy estimates becomes difficult with larger grammars (Roark et al., 2009). Appendix B provides more detail about the grammar.

The full entropy estimates derived from the grammar 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). 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

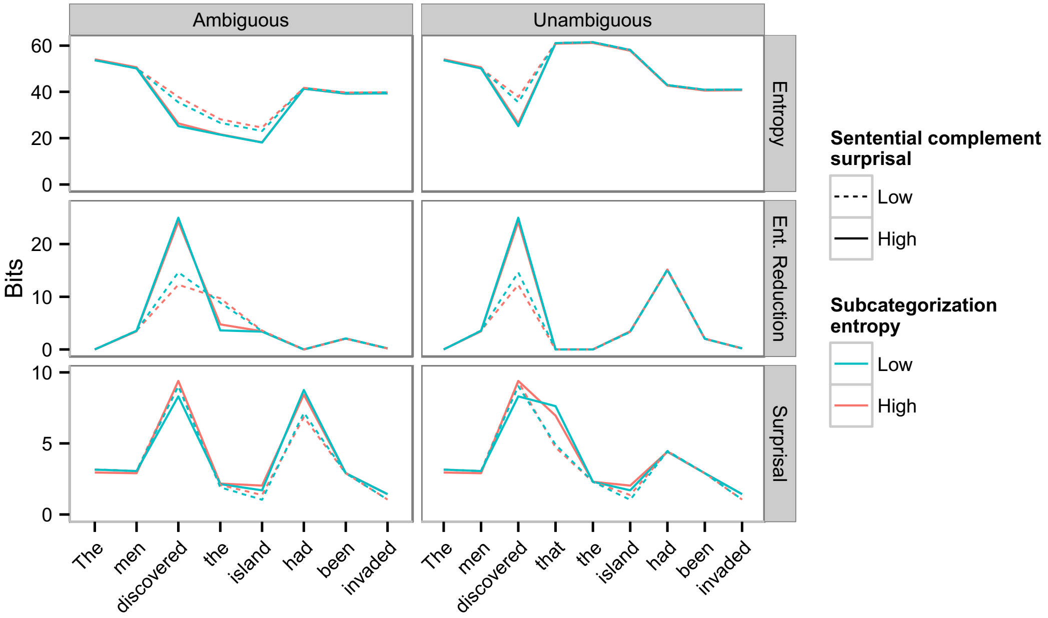

Fig. 5. Word-by-word entropy, entropy reduction, and surprisal predictions derived from the probabilistic context-free grammar based on the Penn Treebank. Predictions are averaged within each of the four conditions of the factorial design. Following Hale (2006), we define entropy reduction at a word to be 0 when entropy after the word is greater than entropy before it.

surprisal estimates derived from the PCFG, averaged within each of the four conditions of the factorial design (high vs. low SC surprisal × 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, based on the PCFG-derived estimates (see Table 3 for an overview). Finally, we analyze the effects of entropy and entropy reduction on RTs. Since entropy and entropy reduction tend to be highly correlated, we assess the effect of each of the variables in a separate model. Instead of the factorial SC 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. All predictors were averaged across all of the words in a region, then centered and standardized before being entered into the model. Unless mentioned otherwise, collinearity did not play a role (correlations between predictors rs < .5).

5.1. Subject region

The same eight main subjects were used in all four conditions of the factorial design (see Fig. 5); the grammar-based predictors therefore did not differ across conditions in

Table 3
Predictions made by uncertainty-based theories (based on full entropy)

	Verb	Ambiguous	Disambiguating
Competition	Lower SC surprisal, Higher subcat entropy → Higher full entropy → Longer RTs ✗	Longer RTs in unambiguous than in ambiguous sentences ✗	
Entropy reduction	Lower SC surprisal, Higher subcat entropy → Higher full entropy → Lower full entropy reduction → Shorter RTs ✓	Shorter RTs in unambiguous than in ambiguous sentences ✓	Longer RTs in unambiguous than ambiguous sentences ⊘
Within ambiguous sentences only:			
Competition		Lower SC surprisal, Higher subcat entropy → Longer RTs ⊘	
Entropy reduction		Lower SC surprisal, Higher subcat entropy → Longer RTs ⊘	

Notes. 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 region). 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 ✗.

this region. The words that made up those eight subjects varied in their parts of speech: *two people* is a numeral followed by a plural noun (CD NNS in Penn Treebank notation), whereas *the man* is a determiner followed by a singular noun (DT NN). This resulted in some limited variability within each condition in the grammar-derived predictors for this region (Fig. 7). There was a strong correlation between entropy and entropy reduction ($r = -.91$).

5.1.1. Results

Linear mixed-effects models did not yield any significant effects in this region, in either the entropy or the entropy reduction analysis (all ps > .3).

5.2. 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 highly inversely correlated in the verb region ($r = -.98$), since entropy before the verb is similar across items (Fig. 7). As is the case in general, the competition hypothesis predicts a positive correlation between RTs entropy, and the entropy reduction hypothesis predicts a positive correlation with entropy reduction.

5.2.1. Results

The entropy reduction analysis found a positive effect of entropy reduction on RTs at the verb ($\beta = 0.014$, $p = .047$). The entropy analysis found a marginal negative effect of entropy ($\beta = -0.012$, $p = .08$). Neither analysis revealed any other effects (*p*s > .5).

The effect of entropy reduction is in the direction predicted by the entropy reduction hypothesis; the direction of the entropy effect is opposite to that predicted by the competition hypothesis. To the extent that RTs on the verb region support either of the two hypotheses, they thus argue in favor of the entropy reduction hypothesis.

RTs on the verb region also shed light on the unexpected numerical difference in RTs between verbs with low and high SC surprisal verbs in the factorial (i.e., single-step entropy) analysis. SC surprisal is one of the major factors that determine full entropy at the verb. This suggests that the apparent effect of SC surprisal in the factorial analysis may be an artifact of its correlation with entropy reduction.

5.3. 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 *that* + determiner increases the probability of an SC to 1, causing full entropy to rise sharply. In

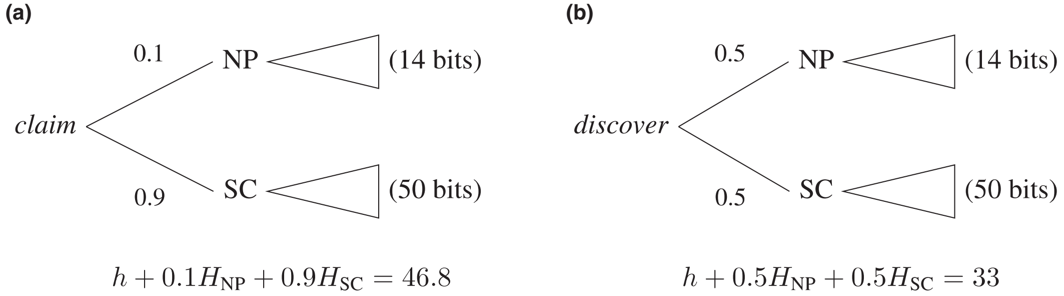
(a)
0.1
NP
(14 bits)
claim
0.9
SC
(50 bits)
$h + 0.1H_{\mathrm{NP}} + 0.9H_{\mathrm{SC}} = 46.8$
(b)
0.5
NP
(14 bits)
discover
0.5
SC
(50 bits)
$h + 0.5H_{\mathrm{NP}} + 0.5H_{\mathrm{SC}} = 33$

Fig. 6. Illustration of the effect of the internal entropy of the verb's complements on full entropy after the verb: (a) in a verb with an SC probability of 0.9; (b) in a verb with an SC probability of 0.5. 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).

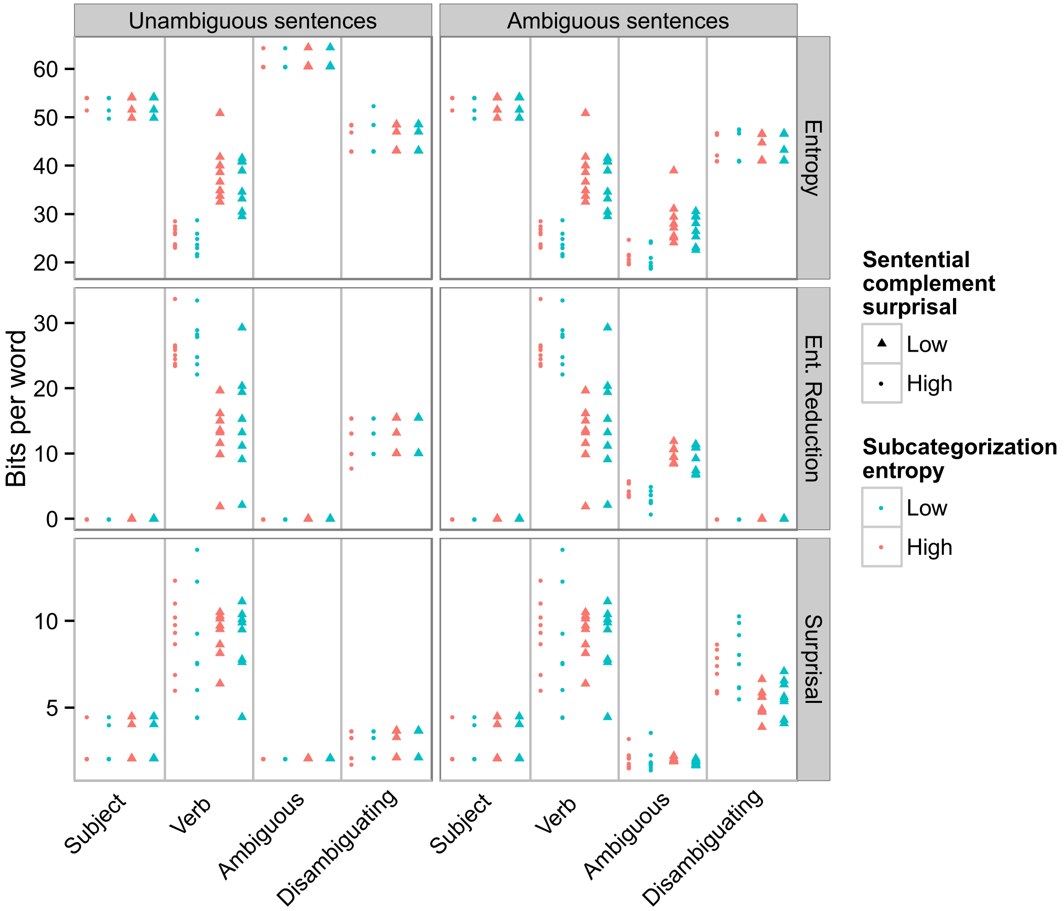

Fig. 7. Variability across items of PCFG-derived variables (entropy, entropy reduction, and surprisal). The values plotted are the mean of each variable across the region. Following Hale (2006), entropy reduction at a word is defined to be 0 when entropy after that word is greater than entropy before it.

ambiguous sentences, on the other hand, some of the probability mass is reserved to the relatively low-entropy NP complement; this results in overall lower entropy. Counterintuitively, then, a strong version of the competition hypothesis, according to which predicted parses lead to competition prior to the presence of bottom-up evidence for those parses, would predict 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 under the entropy reduction hypothesis to affect the 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 RT differences associated with the factorial conditions: 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*) do not depend on the main verb's subcategorization probabilities. The small differences that do exist across individual items (Fig. 7) stem from part of speech differences in the "ambiguous" noun phrase (plural vs. singular noun).

5.3.1. Results

Entropy and entropy reduction were highly correlated in the ambiguous region, though not quite as highly as in the verb regions ($r = -.78$). We first conducted the analyses collapsing across ambiguous and unambiguous sentences. Entropy reduction correlated with increased RTs ($\beta = 0.042$, $p < .001$), and entropy correlated with decreased RTs ($\beta = -0.043$, $p < .001$). No other effects reached significance in either analysis (ps > .5).

These results are predicted by the entropy reduction hypothesis, and are the opposite of the predictions of the competition hypothesis. Since entropy effects in the ambiguous region are highly confounded with Ambiguity (entropy: $r = .99$; entropy reduction: $r = -.85$), however, this result has to be interpreted with caution. For example, it is possible that the observed effect stems from other differences between the ambiguous and unambiguous sentences, such as the presence of temporary ambiguity.

To test whether the effect of the entropy measures was solely carried by Ambiguity, we repeated both analyses over ambiguous sentences only (there was little variability across items in unambiguous sentences; see Fig. 7). None of the predictors in either the entropy reduction model or the entropy model had a significant effect on RTs (all ps > .3). This suggests that the effects found in the first analyses (over ambiguous and unambiguous sentences) were driven by the difference between ambiguous and unambiguous sentences. It leaves open, however, whether the lack of any effects within ambiguous sentences is due to the reduced variability in entropy and entropy reduction within ambiguous sentences as compared to between ambiguous and unambiguous sentences (cf. Fig. 7).

5.4. 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 systematic 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 disambiguating 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 any difference based on the main verb's subcategorization distribution, even within ambiguous sentences: 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 only have an SC parse. Consequently, word-by-word entropy and entropy reduction estimates will only vary with the syntactic categories of the verbal complex in the disambiguating region, for example, *had been invaded* (VBD VBN VBN) versus *should be reported* (MD VB VBD).

Finally, surprisal predicts that the disambiguating region should be read more slowly in ambiguous sentences, where it is in fact disambiguating.

5.4.1. Results

Following our analyses of the ambiguous region, we excluded the Ambiguity factor from the analysis (i.e., collapsed all sentences). Surprisal had a significant effect in both the entropy analysis ($\beta = 0.06$, $p = .01$) and the entropy reduction analysis ($\beta = 0.05$, $p = .03$). None of the other predictors was significant (*p*s > .1).

5.5. Summary of full entropy analyses

A summary of the predictions of the competition and entropy reduction hypotheses and the results of the full entropy analyses is given in Table 3. We found a significant effect of entropy reduction on RTs at the verb: Higher entropy reduction at the verb correlated with longer RTs. The interpretation of this effect is somewhat complicated by the fact that we also observed a marginal effect of entropy in the same region. The high correlation between entropy and entropy reduction makes it impossible to distinguish these two effects via model comparison. However, the direction of the observed pattern is consistent with the predictions of the entropy reduction hypothesis and inconsistent with the predictions of the competition hypothesis. Similarly, RTs at the ambiguous region support the predictions of the entropy reduction hypothesis but are inconsistent with the predictions of the competition hypothesis. RTs on the verb and ambiguous region thus provide support for the entropy reduction hypothesis over the competition hypothesis, while comparing both hypotheses on equal ground (using the same representational assumptions, the same stimuli, and the same control factors).

It is important to keep in mind that syntactic structure is kept constant only at the verb; when both Ambiguity conditions are included in the analysis of the ambiguous region, the regression model collapses across different syntactic structures (much like the comparison between subject and object relative clauses in Yun et al., 2015). Consequently, the longer RTs in the ambiguous region of ambiguous sentences may reflect an unrelated factor, such as 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). The results at the verb therefore constitute stronger support for the role of uncertainty than the results at the ambiguous region.

Reading times at the disambiguating region were predicted by neither the competition nor the entropy reduction hypothesis; the only significant predictor of RTs in this region was surprisal. This result has to be interpreted with caution, however: Surprisal was highly correlated with both entropy ($r = -.79$) and entropy reduction ($r = -.71$). This means that surprisal and entropy reduction effects are predicted to operate in the opposite direction from each other. It is thus possible that the strong effects of surprisal masks any effect of entropy reduction.

6. General discussion

Building on recent evidence that readers maintain expectations over upcoming syntactic structure, this study has investigated how readers' parsing performance is affected by the probability distribution of those expectations, focusing specifically on uncertainty about upcoming structure. We outlined two hypotheses about the potential role of uncertainty in parsing: the competition hypothesis, according to which 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); and the entropy reduction hypothesis, according to which processing is slowed down by any word that *reduces* uncertainty (Hale, 2006).

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. Much previous work has employed single-step or other nonfull estimates of entropy (e.g., Frank, 2013; Roark et al., 2009; Wu et al., 2010; but see Hale, 2003, 2006). This is potentially problematic as the entropy reduction hypothesis is formulated in full entropy terms (Hale, 2006).

Indeed, the distinction between single-step and full entropy turned out to be critical for the current study: Single-step entropy did not affect RTs (aside for an unexpected and likely spurious interaction with surprisal in the ambiguous region), 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 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 not confirmed applies to the disambiguating region. Here, the entropy reduction hypothesis predicts that ambiguous sentences should be read faster than unambiguous ones. The opposite was observed. As outlined above, however, this does not necessarily provide a strong argument against the entropy reduction hypothesis: In the disambiguating region, surprisal is expected to have the opposite effect from entropy reduction. Given the strong surprisal effects at the disambiguating region, all that can be concluded from this result is that (in this case) surprisal may have a stronger effect on processing difficulty than entropy reduction. Indeed, while both surprisal and entropy reduction had statistically significant effects on RTs—at the disambiguating region and at the verb, respectively—the size of the surprisal effect was larger than the size of the entropy reduction effect ($\beta = 0.06$ vs. $\beta = 0.014$).

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

What is the lookahead distance n that humans use in sentence processing? 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 RTs can be predicted by full but not single-step entropy 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 the amount of lookahead increased the extent to which entropy reduction predicted RTs.[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. Green & 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 *slowly* in unambiguous sentences than in ambiguous ones—contrary to our findings (which replicate Kennison, 2001; Pickering & Traxler, 1998). 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.

Entropy estimates and the definition of what constitutes a single derivation step may also 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). 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), have been shown to 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 steps 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.

7. 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). The design of the current study addressed differences between previous works that complicated an evaluation of the entropy reduction and competing hypotheses (cf. Levy & Gibson, 2013). However, the entropy reduction hypothesis failed to predict RTs where it was in conflict with the surprisal hypothesis

(Hale, 2001; Levy, 2008). This suggests that predictability (surprisal) and uncertainty both play a role in explaining processing difficulty in sentence processing. Modeling of the RT results further suggests that 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.

Acknowledgments

A previous version of this work was presented at the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics. We thank Vera Demberg, Stefan Frank, Alec Marantz, and Brian Roark for discussion, John Hale for sharing his code with us, and Andrew Watts for technical assistance with the experiment. This work was supported by an Alfred P. Sloan Fellowship and NSF CAREER (IIS-1150028) to T. Florian Jaeger.

Notes

1. If the subcategorization distribution has only two potential outcomes, SC and another frame (e.g., a direct object), then the surprisal of an SC, given by $-\log_2 p_{SC}$, is deterministically related to the entropy of the distribution, given by $-(p_{SC} \log_2 p_{SC} + (1 - p_{SC}) \log_2 (1 - p_{SC}))$.
2. We only considered active frames; after the verb has been read, passive frames such as *was discovered by* are no longer compatible with the sentence. 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 (four 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 (Fine, Jaeger, Farmer, & Qian, 2013). We investigated whether the effects of surprisal and entropy change over the course of the experiment. We added list position to the region-by-region linear mixed-effects models. There were robust main effects of list position, such that participants became faster overall in later trials across regions (*p*s < .001). Crucially, however, the order effect did not interact with SC surprisal (all *p*s > .1). We conclude that there is no evidence that participants adapted their expectations over the course of the experiment.
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 unavailable 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. Due to model convergence issues, we had to exclude the random by-subject slopes for some of the interactions. Specifically, we excluded the slope for the three-way interaction in all four regions. For the main subject and disambiguating region, we additionally excluded the by-subject slope for the SC-surprisal × Ambiguity interaction.
6. Since the same set of eight NP subjects was used repeatedly in each of the four conditions, the NP subject can be seen as a random effect drawn from a population, and should arguably be modeled as such. We used forward model selection to test whether this random effect was necessary; for each region, we compared a model that included random intercepts and slopes for items and participants only (i.e., the models reported in the text) to a model that additionally included a random intercept and slopes for each predictor in our 2 × 2 × 2 design. In all three regions (verb, ambiguous and disambiguating), adding this random effect did not improve the likelihood of the model significantly (all ps $> .4$).
7. 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.

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

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

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

Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Chen, Z., Hunter, T., Yun, J., & Hale, J. (2014). Modeling sentence processing difficulty with a conditional probability calculator. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1856–1857). Austin, TX: Cognitive Science Society.

Clifton, C., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass*, *2*(2), 234–250.

Collins, M. (1999). *Head-driven statistical models for natural language parsing* (Doctoral dissertation). University of Pennsylvania, Philadelphia, PA.

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

Demberg, V., & Keller, F. (2008). Data from eyetracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Elman, J. L., Hare, M., & McRae, K. (2005). Cues, constraints, and competition in sentence processing. In M. Tomasello, & D. Slobin (Eds.), *Beyond nature-nurture: Essays in honor of Elizabeth Bates* (pp. 111–138). Mahwah, NJ: Erlbaum.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505.

Fine, A. B., & Jaeger, T. F. (2014). The role of verb repetition in cumulative syntactic priming in comprehension. (Manuscript submitted for publication.)

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*(10), e77661.

Francis, W., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Gahl, S., Jurafsky, D., & Roland, D. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods*, *36*(3), 432–443.

Garnsey, S., Pearlmutter, N., Myers, E., & Lotocky, M. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Green, M. J., & Mitchell, D. C. (2006). Absence of real evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *55*(1), 1–17.

Grenander, U. (1967). Syntax-controlled probabilities (Tech. Rep.). Providence, RI: Brown University Division of Applied Mathematics.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Stroudsburg, PA: Association for Computational Linguistics.

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

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jennings, F., Randall, B., & Tyler, L. K. (1997). Graded effects of verb subcategory preferences on parsing: Support for constraint-satisfaction models. *Language and Cognitive Processes*, *12*(4), 485–504.

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, *24*(4), 613–632.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

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

Kallmeyer, L. (2010). *Parsing beyond context-free grammars*. Heidelberg: Springer.

Kennison, S. M. (2001). Limitations on the use of verb information during sentence comprehension. *Psychonomic Bulletin & Review*, *8*, 132–137.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 423–430). Stroudsburg, PA: Association for Computational Linguistics.

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

Levy, R., & Gibson, E. (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, *4*, 229.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Linzen, T., Marantz, A., & Pylkkänen, L. (2013). Syntactic context effects in visual word recognition: An MEG study. *The Mental Lexicon*, *8*(2), 117–139.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

McDonald, S. A., & Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648–652.

McDonald, S. A., & Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Petrov, S., & Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 404–411). Stroudsburg, PA: Association for Computational Linguistics.

Pickering, M. J., & Traxler, M. (1998). Plausibility and the recovery from garden paths: An eyetracking study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 940–961.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*(2), 249–276.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 324–333). Stroudsburg, PA: Association for Computational Linguistics.

Roark, B., & Johnson, M. (1999). Efficient probabilistic top-down and left-corner parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 421–428). Stroudsburg, PA: Association for Computational Linguistics.

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

Spivey-Knowlton, M., & Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, *55*(3), 227–267.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 528–553.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*. doi:10.1093/cercor/bhv075 [Epub ahead of print].

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1189–1198). Stroudsburg, PA: Association for Computational Linguistics.

Yun, J., Chen, Z., Hunter, T., Whitman, J., & Hale, J. (2015). Uncertainty in processing relative clauses across East Asian languages. *Journal of East Asian Linguistics*, *24*(2), 113–148.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Appendix A: List of materials

A.1. 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.

A.2. 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.

A.3. 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.

A.4. 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

B.1. Definitions

A probabilistic context-free grammar (PCFG) consists of a set of nonterminals 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 α is a sequence of terminals or nonterminals (e.g., $VP \rightarrow V\ NP$); and a function ρ 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 $\rho(VP_{break} \rightarrow V_{break}\ NP) = 0.3$ and $\rho(VP_{hit} \rightarrow V_{hit}\ NP) = 0.7$.

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 the following:

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

The full entropy of a nonterminal is defined recursively, as the sum of two terms: the single-step entropy of the nonterminal and the expected sum of the full entropy of any nonterminals that can be derived from the nonterminal. Formally, the full entropy $H(a_i)$ of nonterminal a_i is given by (Grenander, 1967):

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

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 as follows:

$$H = (I - A)^{-1} h,$$

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.

B.2. Full entropy estimation

Word-by-word entropy estimates for our materials were derived using the Cornell Conditional Probability Calculator (Chen, Hunter, Yun, & Hale, 2014) 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 of the grammar by removing rules that included punctuation, rules that occurred less than 100 times (out of the total 1,320,490 nonterminal productions) and rules that had a probability of less than .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 Gahl et al., 2004). 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 $VP_{discover} \rightarrow V_{discover}\, NP$ was defined to be:

$$\rho(VP_{discover} \rightarrow V_{discover}\, NP) = \frac{1}{2}\frac{freq(discover)P(NP|discover)}{\sum_i freq(v_i)}$$

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 (Collins, 1999) or by adding contextual information to some of the tags; for example, the NP tag may be split into NP^VP for an NP whose parent is a 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; similarly, for auxiliary verbs we replaced VBD with VBD*[had]* and VBN with VBN*[been]*. 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, for example, 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.

The focus of this study is on *syntactic* surprisal and entropy, that is, on the portion of those measures that is due to the part of speech of the word rather than its identity (Roark et al., 2009). To tease out the syntactic component of these measures, then, the input to the parser consisted of parts of speech, with the exception of the main verb; for example:

(3)	The	men	discovered	the	island	had	been
	DT	NNS	discovered	DT	NN	*VBD[had]*	*VBN[been]*
	invaded	by	the	enemy.			
	VBN	IN	DT	NN			

Appendix C: Word-by-word reading times

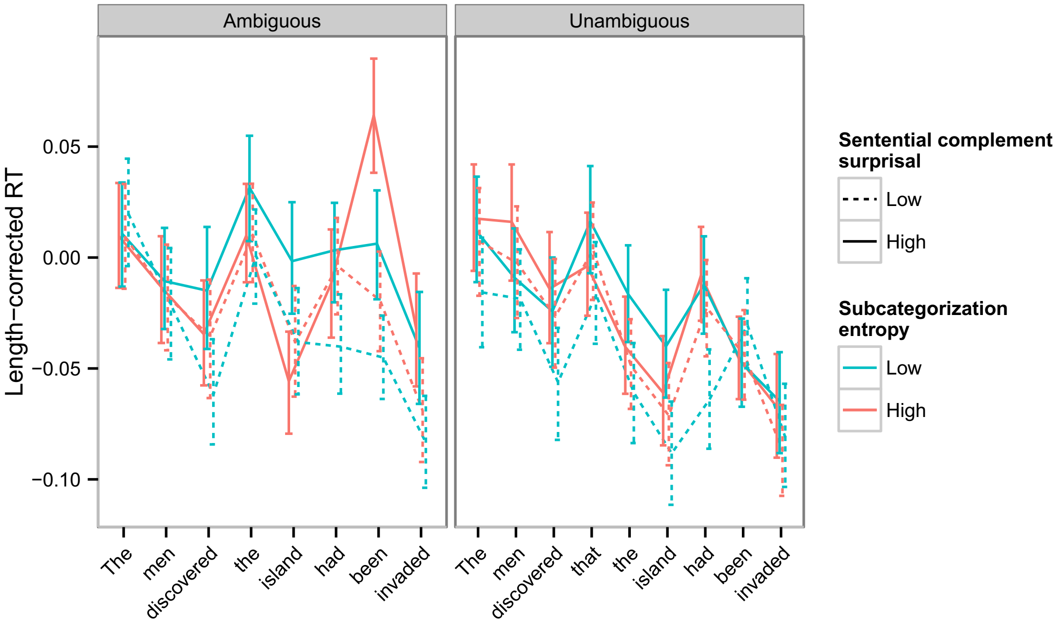

Fig. 8. Word-by-word reading times averaged by the conditions of the factorial design. Error bars represent bootstrapped 95% confidence intervals.