Investigating the role of entropy in sentence processing

Tal Linzen
Department of Linguistics
New York University
linzen@nyu.edu

T. Florian Jaeger
Brain and Cognitive Sciences
University of Rochester
fjaeger@bcs.rochester.edu

Abstract

We outline four ways in which uncertainty might affect comprehension difficulty in human sentence processing. These four hypotheses motivate a self-paced reading experiment, in which we used verb subcategorization distributions to manipulate the uncertainty over the next step in the syntactic derivation (single step entropy) and the surprisal of the verb's complement. We additionally estimate word-by-word surprisal and total entropy over parses of the sentence using a probabilistic context-free grammar (PCFG). Surprisal and total entropy, but not single step entropy, were significant predictors of reading times in different parts of the sentence. This suggests that a complete model of sentence processing should incorporate both entropy and surprisal.

1 Introduction

Predictable linguistic elements are processed faster than unpredictable ones. Specifically, processing load on an element A in context C is linearly correlated with its surprisal, $-\log_2 P(A|C)$ (Smith and Levy, 2013). This suggests that readers maintain expectations as to the upcoming elements: likely elements are accessed or constructed in advance of being read. While there is substantial amount of work on the effect of predictability on processing difficulty, the role (if any) of the distribution over expectations is less well understood.

Surprisal predicts that the distribution over competing predicted elements should not affect reading times: if the conditional probability of a word A is $P(A|C)$, reading times on the word will be proportional to $-\log_2 P(A|C)$, regardless of whether the remaining probability mass is distributed among two or a hundred options.

The **entropy reduction hypothesis** (Hale, 2003; Hale, 2006), on the other hand, accords a central role to the distribution over predicted parses. According to this hypothesis, an incoming element is costly to process when it entails a change from a state of high uncertainty (e.g., multiple equiprobable parses) to a state of low uncertainty (e.g., one where a single parse is much more likely than the others). Uncertainty is quantified as the entropy of the distribution over complete parses of the sentence; that is, if A^i is the set of all possible parses of the sentence after word w_i, then the uncertainty following w_i is given by

$$H_{w_i} = -\sum_{a \in A^i} P(a) \log_2 P(a) \qquad (1)$$

Processing load in this hypothesis is proportional to the entropy reduction caused by w_n:[1]

$$\text{ER}(w_n) = \max\{H_{w_{n-1}} - H_{w_n}, 0\} \qquad (2)$$

A third hypothesis, which we term the **competition hypothesis**, predicts that higher competition among potential outcomes should result in increased processing load at the point at which the competing parses are still valid (McRae et al., 1998; Tabor and Tanenhaus, 1999). This contrasts with the entropy reduction hypothesis, according to which processing cost arises when competition is *resolved*. Intuitively, the two hypotheses make inversely correlated predictions: on average, there will be less competition following words that reduce entropy. A recent study found that reading times on w_i correlated positively with entropy following w_i, providing support for this hypothesis (Roark et al., 2009).

The fourth hypothesis we consider, which we term the **commitment hypothesis**, is derived from

[1] No processing load is predicted for words that increase uncertainty.

Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics, pages 10–18,
Baltimore, Maryland USA, June 26 2014.

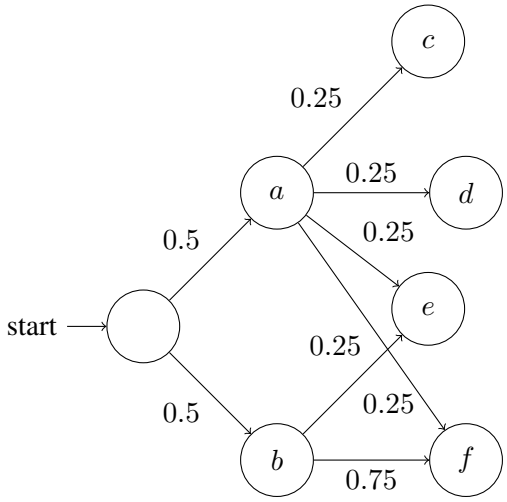

Figure 1: Example language. Output strings are indicated inside the nodes, and transition probabilities are indicated on the edges. For example, the probability of the sentence bf is 0.5×0.75.

the event-related potential (ERP) literature on contextual constraint. Studies in this tradition have compared the responses to a low-predictability word across two types of context: high-constraint contexts, in which there is a strong expectation for a (different) word, and low-constraint ones, which are not strongly predictive of any individual word. There is increasing evidence for an ERP component that responds to violations of a strong prediction (Federmeier, 2007; Van Petten and Luka, 2012). This component can be interpreted as reflecting disproportional commitment to high probability predictions at the expense of lower probability ones, a more extreme version of the proposal that low-probability parses are pruned in the presence of a high-probability parse (Jurafsky, 1996). Surprisal is therefore expected to have a larger effect in high constraint contexts, in which entropy was low before the word being read. Commitment to a high probability prediction may also result in increased processing load at the point at which the commitment is made.

We illustrate these four hypotheses using the simple language sketched in Figure 1. Consider the predictions made by the four hypotheses for the sentences ae and be. Surprisal predicts no difference in reading times between these sentences, since the conditional probabilities of the words in the two sentences are identical (0.5 and 0.25 respectively).

The competition hypothesis predicts increased reading times on the first word in ae compared to be, because the entropy following a is higher than the entropy following b (2 bits compared to 0.71). Since all sentences in the language are two word long, entropy goes down to 0 after the second word in both sentences. This hypothesis therefore does not predict a reading time difference on the second word e.

Moving on to the entropy reduction hypothesis, five of the six possible sentences in the language have probability 0.5×0.25, and the sixth one (bf) has probability 0.5×0.75. The full entropy of the grammar is therefore 2.4 bits. The first word reduces entropy in both ae and be (to 2 and 0.71 bits respectively), but entropy reduction is higher when the first word is b. The entropy reduction hypothesis therefore predicts longer reading times on the first word in be than in ae. Conversely, since entropy goes down to 0 in both cases, but from 2 bits in ae compared to 0.71 bits in be, this hypothesis predicts longer reading times on e in ae than in be.

Finally, the commitment hypothesis predicts that after b the reader will become committed to the prediction that the second word will be f. This will lead to longer reading times on e in be than in ae, despite the fact that its conditional probability is identical in both cases. If commitment to a prediction entails additional work, this hypothesis predicts longer reading times on the first word when it is b.

This paper presents an reading time study that aims to test these hypotheses. Empirical tests of computational theories of sentence processing have employed either reading time corpora (Demberg and Keller, 2008) or controlled experimental materials (Yun et al., 2010). The current paper adopts the latter approach, trading off a decrease in lexical and syntactic heterogeneity for increased control. This paper is divided into two parts. Section 2 describes a reading time experiment, which tested the predictions of the surprisal, competition and commitment hypotheses, as applied to the entropy over the next single step in the syntactic derivation.[2] We then calculate the total entropy (up to an unbounded number of derivation steps) at each word using a PCFG; Section 3 describes how this grammar was constructed, overviews the predictions that it yielded in light of the four hypotheses, and evaluates these predictions on the results of the reading time experiment.

[2]We do not test the predictions of the entropy reduction hypothesis in this part of the paper, since that theory explicitly only applies to total rather than single-step entropy.

2 Reading time experiment

2.1 Design

To keep syntactic structure constant while manipulating surprisal and entropy over the next derivation step, we took advantage of the fact that verbs vary in the probability distribution of their syntactic complements (subcategorization frames). Several studies have demonstrated that readers are sensitive to subcategorization probabilities (Trueswell et al., 1993; Garnsey et al., 1997).

The structure of the experimental materials is shown in Table 1. In a 2x2x2 factorial design, we crossed the surprisal of a sentential complement (SC) given the verb, the entropy of the verb's subcategorization distribution, and the presence or absence of the complementizer *that*. When the complementizer is absent, the region *the island* is ambiguous between a direct object and an embedded subject.

Surprisal theory predicts an effect of SC surprisal on the disambiguating region in ambiguous sentences (sentences without *that*), as obtained in previous studies (Garnsey et al., 1997), and an effect of SC surprisal on the complementizer *that* in unambiguous sentences. Reading times should not differ at the verb: in the minimal context we used (*the men*), the surprisal of the verb should be closely approximated by its lexical frequency, which was matched across conditions.

The competition hypothesis predicts a positive main effect of subcategorization frame entropy (subcategorization frame entropy) at the verb: higher uncertainty over the syntactic category of the complement should result in slower reading times.

The commitment hypothesis predicts that the effect of surprisal in the disambiguating region should be amplified when subcategorization frame entropy is low, since the readers will have committed to the competing high probability frame. If the commitment step in itself incurs a processing cost, there should be a negative main effect of subcategorization frame entropy at the verb.

This experimental design varies the entropy over the single next derivation step: it assumes that the parser only predicts the identity of the subcategorization frame, but not its internal structure. Since the predictions of the entropy reduction hypothesis crucially depend on predicting the internal structure as well, we defer the discussion of that hypothesis until Section 3.

The men	discovered	(that)	the island	had been invaded	by the enemy.
mat. subj.	*verb*	*that*	*emb. subj.*	*emb. verb complex*	*rest*

Table 1: Structure of experimental materials (mat. = matrix, emb. = embedded, subj. = subject).

2.2 Methods

2.2.1 Participants

128 participants were recruited through Amazon Mechanical Turk and were paid $1.75 for their participation.

2.2.2 Materials

32 verbs were selected from the Gahl et al. (2004) subcategorization frequency database, in 4 conditions: high vs. low SC surprisal and high vs. low subcategorization frame entropy (see Table 2). Verbs were matched across conditions for length in characters and for frequency in SUBTLEX-US corpus (Brysbaert and New, 2009). A sentence was created for each verb, following the structure in Table 1. Each sentence had two versions: one with the complementizer *that* after the verb and one without it. The matrix subjects were minimally informative two-word NPs (e.g. *the men*). Following the complementizer (or the verb, if the complementizer was omitted) was a definite NP (*the island*), which was always a plausible direct object of the matrix verb.

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

In addition to the target sentences, the experiment contained 64 filler sentences, with various complex syntactic structures.

2.2.3 Procedure

The sentences were presented word by word in a self-paced moving window paradigm. The participants were presented with a Y/N comprehension question after each trial. The participants did not

	NP	Inf	PP	SC	SC s.	SFE
forget	0.55	0.14	0.2	0.09	3.46	1.7
hear	0.72	0	0.17	0.11	3.22	1.12
claim	0.36	0.12	0	0.45	1.15	1.71
sense	0.61	0	0.02	0.34	1.55	1.18

Table 2: A example verb from each of the four conditions. On the left, probabilities of complement types: noun phrase (NP), infinitive (Inf), prepositional phrase (PP), sentential complement (SC); on the right, SC surprisal and subcategorization frame entropy.

receive feedback on their responses. The experiment was conducted online using a Flash application written by Harry Tily (now at Nuance Communications).

2.2.4 Statistical analysis

Subjects were excluded if their answer accuracy was lower than 75% (two subjects), or if their mean reading time (RT) differed by more than 2.5 standard deviations from the overall mean RT across subjects (two subjects). The results reported in what follows are based on the remaining 124 subjects (97%).

We followed standard preprocessing procedure. Individual words were excluded if their raw RT was less than 100 ms or more than 2000 ms, or if the log-transformed RT was more than 3 standard deviations away from the participant's mean. Log RTs were length-corrected by taking the residuals of a mixed-effects model (Bates et al., 2012) that had log RT as the response variable, word length as a fixed effect, and a by-subject intercept and slope.

The length-corrected reading times were regressed against the predictors of interest, separately for each region. We used a maximal random effect structure. All *p* values for fixed effects were calculated using model comparison with a simpler model with the same random effect structure that did not contain that fixed effect.

2.3 Results

Reading times on the matrix subject (*the men*) or matrix verb (*discovered*) did not vary significantly across conditions.

The embedded subject *the island* was read faster in unambiguous sentences ($p < 0.001$). Reading times on this region were longer when SC surprisal was high ($p = 0.04$). Models fitted to ambiguous and unambiguous sentences separately revealed that the simple effect of SC surprisal on the embedded subject was significant for unambiguous sentences ($p = 0.02$) but not for ambiguous sentences ($p = 0.46$), though the interaction between SC surprisal and ambiguity did not reach significance ($p = 0.22$).

The embedded verb complex (*had been invaded*) was read faster in unambiguous than in ambiguous sentences ($p < 0.001$). Reading times in this region were longer overall in the high SC surprisal condition ($p = 0.03$). As expected, this effect interacted with the presence of *that* ($p = 0.01$): the simple effect of SC surprisal was not significant in unambiguous sentences ($p = 0.28$), but was highly significant in ambiguous ones ($p = 0.007$). We did not find an interaction between SC surprisal and subcategorization frame entropy (of the sort predicted by the commitment hypothesis).

Subcategorization frame entropy did not have a significant effect in any of the regions of the sentence. It was only strictly predicted to have an effect on the matrix verb: longer reading times according to the competition hypothesis, and (possibly) shorter reading times according to the commitment hypothesis. The absence of an subcategorization frame entropy effect provides weak support for the predictions of surprisal theory, according to which entropy should not affect reading times.

3 Deriving predictions from a PCFG

3.1 Calculating entropy

As mentioned above, the entropy of the next derivation step following the current word (which we term *single-step entropy*) is calculated as follows. If a_i is a nonterminal, Π_i is the set of rules rewriting a_i, and p_r is the application probability of rule r, then the single-step entropy of a_i is given by

$$h(a_i) = -\sum_{r \in \Pi_i} p_r \log_2 p_r \qquad (3)$$

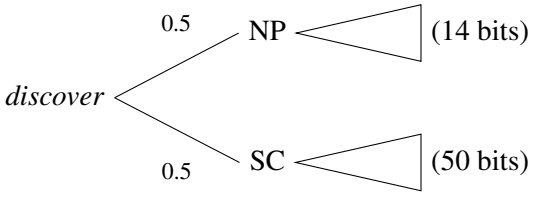

Figure 3: Entropy calculation example: the single step entropy after *discover* is 1 bit; the overall entropy is $1 + 0.5 \times 14 + 0.5 \times 50 = 33$ bits.

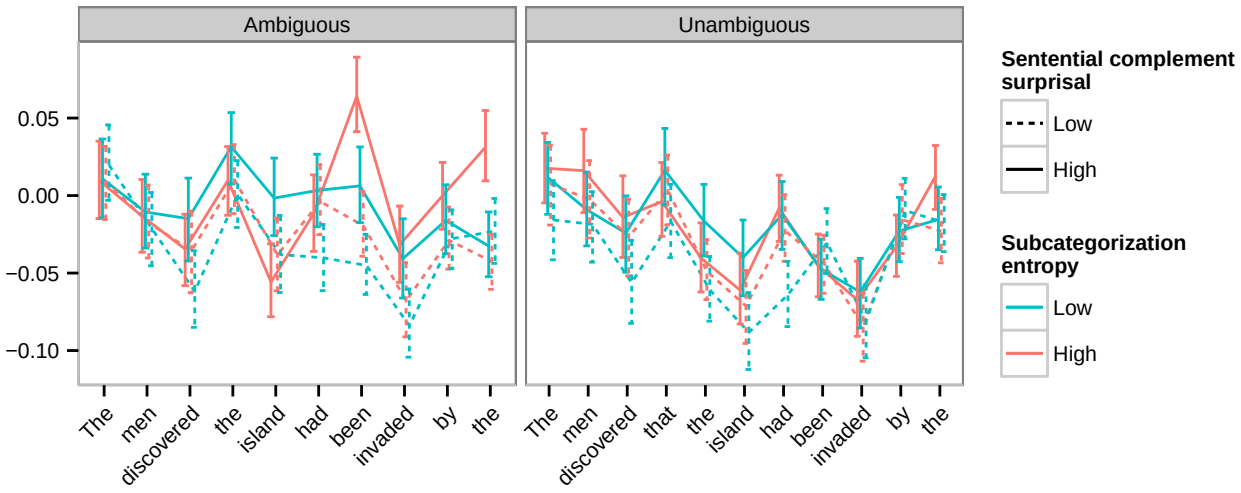

Figure 2: Results of the self-paced reading experiment

The entropy of all derivations starting with a_i (which we term *total entropy*) is then given by the following recurrence:

$$H(a_i) = h(a_i) + \sum_{r \in \Pi_i} p_r \sum_{j=1}^{k_r} H(a_{r,j}) \quad (4)$$

where $a_{r,1}, \ldots, a_{r,k_r}$ are the nonterminals on the right-hand side of r. This recurrence has a closed form solution (Wetherell, 1980; Hale, 2006). The expectation matrix A is a square matrix with N rows and columns, where N is the set of nonterminals. Each element A_{ij} indicates the expected number of times nonterminal a_j will occur when a_i is rewritten using exactly one rule of the grammar. If $h = (h_1, \ldots, h_N)$ is the vector of all single-step entropy values for the N nonterminal types in the grammar, and $H = (H_1, \ldots, H_N)$ is the vector of all total entropy values, then the closed form solution for the recurrence is given by

$$H = (I - A)^{-1} h \quad (5)$$

where I is the identity matrix. The entropy after the first n words of the sentence, H_{w_n}, can be calculated by applying Equation 5 to the grammar formed by intersecting the original grammar with the prefix $w_1, \ldots, w_n$ (i.e., considering only the parses that are compatible with the words encountered so far) (Hale, 2006).

Two points are worth noting about these equations. First, Equation 5 shows that calculating the entropy of a PCFG requires inverting the matrix $I - A$, which is the size of the number of nonterminal symbols in the grammar. This makes it impractical to use a lexicalized grammar, as advocated by Roark et al. (2009), since those grammars have a very large number of nonterminal types.

Second, Equation 4 shows that the entropy of a nonterminal is the sum of its single-step entropy and a weighted average of entropy of the nonterminals it derives. In the context of subcategorization decisions, the number of possible subcategorization frames is small, and the single-step entropy is on the order of magnitude of 1 or 2 bits. The entropy of a typical complement, on the other hand, is much higher (consider all of the possible internal structures that an SC could have). This means that the total entropy H after processing the verb is dominated by the entropy of its potential complements rather than the verb's single-step entropy h (see Figure 3 for an illustration). A lookahead of a single word (as used in Roark et al. (2009)) may therefore be only weakly related to total entropy.

3.2 Constructing the grammar

We used a PCFG induced from the Penn Treebank (Marcus et al., 1993). As mentioned above, the grammar was mostly unlexicalized; however, in order for the predictions to depend on the identity of the verb, the grammar had to contain lexically specific rules for each verb. We discuss these rules at end of this section.

The Penn Treebank tag set is often expanded by adding to each node's tag an annotation of the

node's parent, e.g., marking an NP whose parent is a VP as NP_VP (Klein and Manning, 2003). While systematic parent annotation would have increased the size of the grammar dramatically, we did take the following minimal steps to improve parsing accuracy. First, the word *that* is tagged in the Penn Treebank as a preposition (IN) when it occurs as a subordinating conjunction. This resulted in SCs being erroneously parsed as prepositional phrases. To deal with this issue, we replaced the generic IN with IN[*that*] whenever it referred to *that*.

Second, the parser assigned high probability parses to reduced relative clauses in implausible contexts. We made sure that cases that should not be reduced relative clauses were not parsed as such by splitting the VP category into sub-categories based on the leftmost child of the VP (since only VP[VPN] should be able to be a reduced relative), and by splitting SBAR into SBAR[overt] when the SBAR had an overt complementizer and SBAR[none] when it did not.

Following standard practice, we removed grammatical role information and filler-gap annotations, e.g., NP-SUBJ-2 was treated as NP. To reduce the number of rules in the grammar as much as possible, we removed punctuation and the silent element NONE (used to mark gaps, silent complementizers, etc.), 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 tokens respectively. We then applied horizontal Markovization (Klein and Manning, 2003).

Finally, we added lexically specific rules to capture the verbs' subcategorization preferences, based on the Gahl et al. (2004) subcategorization database. The probability of frame f_j following verb v_i was calculated as:

$$P(\text{VP[VBD]} \rightarrow v_i\ f_j) = \frac{1}{2}\frac{P(v_i)P(f_j|v_i)}{\sum_i P(v_i)} \quad (6)$$

In other words, half of the probability mass of production rules deriving VP[VBD] (VP headed by past tense verbs) was taken away from the unlexicalized rules and assigned to the verb-specific rules. The same procedure was performed for VP[VBN] (VP headed by a past participle, with the exception of the verbs *forgot* and *wrote*, which are not ambiguous between the past and past participle forms. The total probability of all rules deriving VP as a specific verb (e.g., *discovered*) was estimated as the corpus frequency of that verb divided by the total corpus frequency of all 32 verbs used in the experiment, yielding a normalized estimate of the relative frequency of that verb.

3.3 Surprisal, entropy and entropy reduction profiles

Word-by-word surprisal, entropy and entropy reduction values for each item were derived from the equations in Section 3.1 using the Cornell Conditional Probability Calculator (provided by John Hale). Figure 4 shows the predictions averaged by the conditions of the factorial design. Surprisal on the verb is always high because this is the only part of the grammar that encodes lexical identity; surprisal on the verb therefore conflates lexical and syntactic surprisal. Surprisal values on all other words are low, with the exception of the point at which the reader gets the information that the verb's complement is an SC: the embedded verb complex in ambiguous sentences, and the complementizer in unambiguous sentence.

The entropy profile is dominated by the fact that SCs have much higher internal entropy than NPs. As a consequence, entropy after the verb is higher whenever an SC is a more likely subcategorization frame. The entropy after high subcategorization frame entropy verbs is higher than that after low subcategorization frame entropy verbs, though the difference is small in comparison to the effect of SC surprisal. In ambiguous sentences, entropy remains higher for low SC surprisal verbs throughout the ambiguous region. Somewhat counterintuitively, entropy *increases* when the parse is disambiguated in favor of an SC. This is again a consequence of the higher internal entropy of a SC: the entropy of the ambiguity between SC and NP is dwarfed by the internal entropy of a SC. The entropy profile for unambiguous sentences is straightforward: it increases sharply when the reader finds out that the complement is a SC, then decreases gradually as more details are revealed about the internal structure of the SC.

The reading time predictions made by the entropy reduction hypothesis are therefore very different than those made by surprisal theory. On the verb, the entropy reduction hypothesis predicts that high SC surprisal verbs will be read more

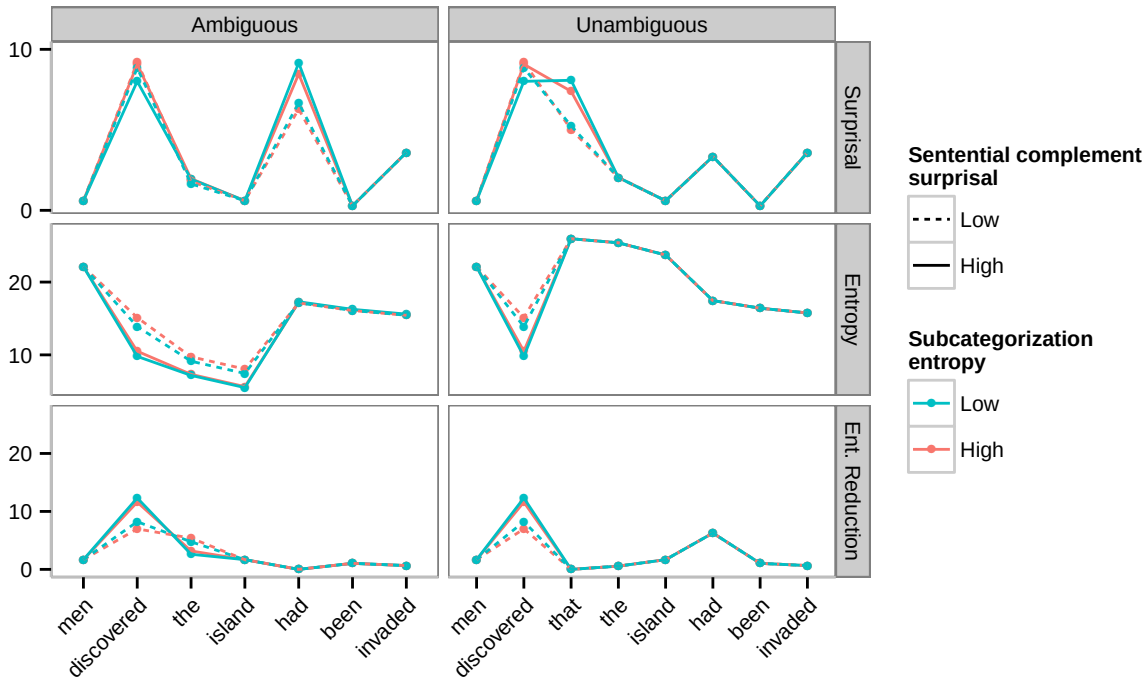

Figure 4: Parser-derived surprisal, entropy and entropy reduction estimates for the stimuli in our experiments, averaged within each condition of the factorial design (first word of sentence and *rest* region excluded).

slowly than low SC surprisal verbs, whereas surprisal predicts no difference. On the disambiguating region in ambiguous sentences, the entropy reduction hypothesis predicts no reading time differences at all, since an increase in entropy is not predicted to affect reading times. In fact, entropy reduction on the word *had* is positive only in unambiguous sentences, so the entropy reduction hypothesis predicts a slowdown in unambiguous compared to ambiguous sentences.

3.4 Evaluation on reading times

We tested whether reading times could be predicted by the word-by-word estimates derived from the PCFG. Since total entropy, entropy reduction and surprisal values did not line up with the factorial design, we used continuous regression instead, again using *lme4* with a maximal random effects structure. We only analyzed words for which the predictions depended on the properties of the verb (as Figure 4 shows, this is only the case for a minority of the words). As outcome variables, we considered both reading times on the word w_i, and a spillover variable computed as the sum of the reading times on w_i and the next word w_{i+1}. The predictors were standardized (separately for each word) to facilitate effect comparison.

Parser-derived entropy reduction values varied the most on the main verb. Since the word following the verb differs between the ambiguous and unambiguous conditions, we added a categorical control variable for sentence ambiguity. In the resulting model, lower entropy (or equivalently, higher entropy reduction values), caused an increase in reading times (no spillover: $\hat{\beta} = 0.014$, $p = 0.05$; one word spillover: $\hat{\beta} = 0.022$, $p = 0.04$). Our design does not enable us to determine whether the effect of entropy on the verb is due to entropy *reduction* or simply entropy. The commitment hypothesis is therefore equally supported by this pattern as is the entropy reduction hypothesis.

The only other word on which entropy reduction values varied across verbs was the first word *the* of the ambiguous region. Neither entropy reduction nor surprisal were significant predictors of reading times on this word.

There was also some variation across verbs in entropy (though not entropy reduction) on the second word of the embedded subject (*island*) in ambiguous sentences; however, entropy was not a significant predictor of reading times on that word. In general, entropy is much higher in the embed-

ded subject region in unambiguous than ambiguous sentences, since it is already known that the complement is an SC, and the entropy of an SC is higher. Yet as mentioned above, reading times on the embedded subject were higher when it was ambiguous ($p < 0.001$).

Finally, PCFG-based surprisal was a significant predictor of reading times on the disambiguating word in ambiguous sentences (no spillover: *n.s.*; one word spillover: $\hat{\beta} = 0.037$, $p = 0.02$; two-word spillover: $\hat{\beta} = 0.058$, $p = 0.001$). In contrast with simple SC surprisal (see Section 2.2.4), PCFG-based surprisal was not a significant predictor of reading times on the complementizer *that* in unambiguous sentences.

4 Discussion

We presented four hypotheses as to the role of entropy in syntactic processing, and evaluated them on the results of a reading time study. We did not find significant effects of subcategorization frame entropy, which is the entropy over the next derivation step following the verb. Entropy over complete derivations, on the other hand, was a significant predictor of reading time on the verb. The effect went in the direction predicted by the entropy reduction and commitment hypotheses, and opposite to that predicted by the competition hypothesis: reading times were higher when post-verb entropy was lower.

Reading times on the embedded subject in ambiguous sentences were increased compared to unambiguous sentences. This can be seen as supporting the competition hypothesis: the SC and NP parses both need to be maintained, which increases processing cost. Yet the parser predictions showed that total entropy on the embedded subject was higher in unambiguous than ambiguous sentences, since the probability of the high-entropy sentential complement is 1 in unambiguous sentences. In this case, then, total entropy, which entails searching enormous amounts of predicted structure, may not be the right measure, and single-step (or n-step) entropy may be a better predictor.

In related work, Frank (2013) tested a version of the entropy reduction hypothesis whereby entropy reduction was not bounded by 0 (was allowed to take negative values). A Simple Recurrent Network was used to predict the next four words in the sentence; the uncertainty following the current word was estimated as the entropy of this quadrigram distribution. Higher (modified) entropy reduction resulted in increased reading times. These results are not directly comparable to the present results, however. Frank (2013) tested a theory that takes into account both positive and negative entropy changes. In addition, a four-word lookahead may not capture the dramatic difference in internal entropy between SCs and NPs, which is responsible for the differential reading times predicted on the matrix. This caveat applies even more strongly to the one-word lookahead in Roark et al. (2009).

In contrast with much previous work, we calculated total entropy using a realistic PCFG acquired from a Treebank corpus. In future work, this method can be used to investigate the effect of entropy in a naturalistic reading time corpus. It will be important to explore the extent to which the reading time predictions derived from the grammar are affected by representational decisions (e.g., the parent annotations we used in Section 3.2). This applies in particular to entropy, which is sensitive to the distribution over syntactic parses active at the word; surprisal depends only the conditional probability assigned to the word by the grammar, irrespective of the number and distribution over the parses that predict the current word, and is therefore somewhat less sensitive to representational assumptions.

5 Conclusion

This paper described four hypotheses regarding the role of uncertainty in sentence processing. A reading time study replicated a known effect of surprisal, and found a previously undocumented effect of entropy. Entropy predicted reading times only when it was calculated over complete derivations of the sentence, and not when it was calculated over the single next derivation step. Our results suggest that a full theory of sentence processing would need to take both surprisal and uncertainty into account.

Acknowledgments

We thank Alec Marantz for discussion and Andrew Watts for technical assistance. This work was supported by an Alfred P. Sloan Fellowship to T. Florian Jaeger.

References

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

M. Brysbaert and B. New. 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.

V. Demberg and F. Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

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

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

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

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

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

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

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

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

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

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

B. Roark, A. Bachrach, C. Cardenas, and C. Pallier. 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: Volume 1-Volume 1*, pages 324–333. Association for Computational Linguistics.

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

W. Tabor and M. K. Tanenhaus. 1999. Dynamical models of sentence processing. *Cognitive Science*, 23(4):491–515.

J. Trueswell, M. Tanenhaus, and C. Kello. 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.

C. Van Petten and B. Luka. 2012. Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2):176–190.

C. S. Wetherell. 1980. Probabilistic languages: A review and some open questions. *ACM Computing Surveys (CSUR)*, 12(4):361–379.

J. Yun, J. Whitman, and J. Hale. 2010. Subject-object asymmetries in Korean sentence comprehension. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.