# PROCESSING EFFECTS OF THE EXPECTATION OF INFORMATIVITY

Richard Futrell

May 2012

(Daniel Jurafsky) Principal Adviser

(Hannah Rohde)

Approved for the University Committee on Graduate Studies

I'd like to give heartfelt thanks to my advisor Dan Jurafsky and my collaborator Hannah Rohde, and also to Michael Ramscar, Tom Wasow, Roger Levy, Ivan Sag, and Gabe Doyle for discussion and encouragement. This work was hugely facilitated by the work of Alex Drummond, who made Ibex and provided hosting for my experiments, and of Klinton Bicknell, who wrote the scripts used to analyze self-paced reading results. Funding for sentence completion tasks came from the Stanford Humanities Center, and funding for self-paced reading studies came from an Andrew Mellon Postdoctoral Fellowship to Hannah Rohde.

# Contents

# List of Tables

# List of Figures

# 1 Predictability in Language

Predictability has emerged as a pivotal factor in human language processing and in cognition more generally. Humans constantly form expectations about what might occur next in their environment; this also occurs during the process of understanding some linguistic material. The outcome of successful learning is to minimize surprise and make aspects of the world maximally predictable, and this applies to language as well. Empirically, the predictability of linguistic material is an excellent predictor of how easily that material is understood.

Predictability in language, however, poses a conundrum because language is used for the communication of information, and information is by definition unpredictable material (Shannon, 1948). As we comprehend language, we have an expectation of informativity, which means that we must expect, potentially paradoxically, that the language producer's utterance will be unpredictable in some way. The intuition is that if the utterance is predictable, then it would be surprising that the producer took the effort to utter it at all.

The aim of this thesis is to investigate empirically the nature of language users' expectation of informativity. I defend the hypothesis that language users are sensitive to cues about the informational topography of linguistic material. That is, language users have a variable expectation of informativity: at some times they expect high information, and at other times they expect low information, allocating processing resources accordingly. This expectation can be modulated by discourse factors and by conventional linguistic structures.

Empirically, I hypothesize that the expectation of informativity can lead to situations where linguistic material that is too predictable leads to processing difficulties. That is, a situation might arise where predictable material is harder to process than unpredictable material. More generally, when language users expect informativity, informative material should be easier to process, while uninformative material should be harder to process because it clashes with higher-level discourse expectations. The predicted effect on processing difficulty is thus a negative interaction between the expectation of linguistic informativity and the actual informativity of the linguistic signal.

I carry out three self-paced reading time studies and one sentence completion study in order to address these questions. The studies were distributed over Amazon's Mechanical Turk, a marketplace where workers complete small tasks for small sums of money, allowing for fast data collection from a large number of unique subjects. The thesis also includes two validation studies replicating classic psycholinguistic effects with Mechanical Turk workers as subjects in order to verify the suitability of Mechanical Turk for carrying out self-paced reading studies.

The thesis is organized as follows:

**Section 1.** A brief review of the empirical and theoretic roles of predictability in human cognition and human language,

**Section 2.** Motivation and exposition of specific hypotheses about the expectation of informativity,

**Section 3.** Description of the use of Mechanical Turk for gathering self-paced reading time data, along with two validation studies for this methodology,

**Section 4.** Analysis of one sentence-completion experiment and three self-paced reading time experiments,

**Section 5.** Conclusions and directions for future studies.

## 1.1 Prediction in online language processing

Prediction and the effects of predictability are pervasive phenomena in human language processing. Behavioral and brain-based studies reveal that, in the process of comprehending language, humans make use of as much available information as they can to predict the speaker's intentions (Tomasello, 2003), a process which often involves predicting upcoming linguistic material. The result of this process is that words that are predictable are interpreted and produced more quickly. I review literature to this effect below.

The formation of predictions is most clearly illustrated in results from the visual world paradigm, in which an experimental subject listens to a language recording while looking at a screen containing multiple objects. Eye tracking is used to determine where on the screen the subject is fixating. Subjects fixate on objects that are relevant to the message being conveyed; for instance, if they hear the sentence *the cat sat on the mat*, they will look toward an illustration of a cat on a mat. Altmann and Kamide (1999) showed that subjects look toward objects that have not yet been mentioned, but which are predictable given what *has* been said so far. Given the sentence fragment *The boy will eat...*, subjects look at an illustration of cake rather than an illustration of a ball. After hearing *The little girl will ride...*, subjects already look toward a picture of a merry-go-round and not a motorcycle (Kamide et al., 2003a). The effect is not limited to predicting the objects of verbs; subjects of verbs, when they follow the verb, are also subject to prediction in German (Kamide et al., 2003b). It seems that the kinds of linguistic material that hearers use to make predictions about upcoming message content is not limited to certain grammatical relations, though issues of scope might still restrict what information is used in the formation of predictions.

When a prediction is in error, surprise and confusion result; these responses provide a window into what exactly is predicted. One index of this prediction error is the N400, an event-related potential in the brain which can be measured using electroencephalography (EEG). It consists of a negative-going deflection peaking about 400 ms after a stimulus, and has been found to arise in response to prediction error in meaningful stimuli (see Kutas and Federmeier (2009) for a review).

N400 response potentials reveal that language comprehenders predict both the form and meaning of upcoming linguistic material. DeLong et al. (2005) set up contexts which bias a comprehender to expect nouns that begin either with a consonant or with a vowel. They observe N400 responses for the indefinite article *an* in a context that favors consonants and for the article *a* in contexts that favor vowels.

Language users predict not only wordforms, but also aspects of the meaning of linguistic material as well. As the predictability of a word in context decreases, the size of the N400 potential increases (Kutas and Hillyard, 1984). If a word is not found to be predictable in a given context using the cloze task or corpus probabilities, that word may still elicit a small N400 potential if it is semantically related to words that have been found to be predictable, thus indicating that language comprehenders are predicting aspects of the meaning of a message and not merely wordforms. For instance, comprehenders have likely never heard the sentence *John cut steak with a cutlass*, but will have a smaller N400 for *cutlass* in this sentence than for *eraser*, because *cutlass* is semantically related to *knife*. These results are consistent with a model in which comprehenders employ semantic similarity-based smoothing to assign high probability to events which they have never observed but which are similar to events they have observed (Dagan et al., 1999; Yarlett, 2008). The prediction of the phonological form of upcoming material may be an intermediate stage in the prediction of the speaker's intentions, or it may be that hearers are predicting both meaning and form in parallel as part of a more general prediction process.

Hearers form expectations in response to more general discourse context in addition to grammatically related words within sentences. In an ERP study, Nieuwland and van Berkum (2006) present hearers with discourses where global context establishes that a certain peanut is an animate character by describing it as dancing. This overrides hearers' prior expectations about the use of the word *peanut* in following sentences and leads to a lack of surprisal effects when the peanut is described as *in love*. In this case, global discourse context shifts hearers' expectations about words referring to the peanut from words describing inanimate actions to words describing animate ones.

The role of probabilistic predictions extends well beyond psycholinguistics. The N400 response to prediction error appears also for stimuli involving music (Daltrozzo and Schön, 2009) and videos (Sitnikova et al., 2003), indicating that predictions are formed for these domains as well. Neural learning in the hippocampus and cortical synapses has been argued to be predictive, rather than correlational, in that the training signal is based on the prediction of upcoming events (Dayan, 2002). The release of dopamine, a key reward signal for reinforcement learning, has found to correlate with prediction error, in that the reward only exists–and learning only happens–when there is a discrepancy between what is predicted and what occurs (Hollerman and

Schultz, 1998; Schultz, 2007).

The general role of predictions and of errors in predictions are formalized in influential models of learning. The prominent Rescorla-Wagner learning rule, as well as the delta rule used for learning in many connectionist networks, learns explicitly from error in prediction (Rescorla and Wagner, 1972; Rescorla, 1988; Siegel and Allan, 1996). Models of representation formation based on the prediction of linguistic labels in their temporal order have led to intriguing findings in the learning of perfect pitch, number words, and categorization in general (Ramscar et al., 2010, 2011b,a). In these studies, the temporal order of labels affects whether humans learn a veridical representation of a concept or a distorted representation optimized for the prediction of one label rather than another.

## 1.2 Predictability eases processing

Behavioral measures such as reaction time also reveal the fulfillment or frustration of comprehenders' expectations. In the case of reading, the predictability of words correlates with shorter reading times (Kliegl et al., 2004) and thus, presumably, easier processing. This subsection focuses specifically on the effects of linguistic predictability rather than the role of predictions in general. To do so, I first review how predictability and uncertainty can be quantified, then review the literature on the processing effects of linguistic predictability.

### 1.2.1 Information-Theoretic Measures

The tools for quantifying predictability come from information theory, where information is identified with unpredictability (Shannon, 1948). If an event is predictable, then the information it conveys is redundant; if the event is unpredictable, it is highly informative. Mathematically, the self-information of an event $e$ is calculated as $-log_2 p(e)$ bits, a number which is large for low probabilities and small for high probabilities.

The informativity of linguistic events can be measured from probabilities derived from corpus or cloze frequency. Consider the probability that an observed word is *house*, written $p(house)$, and the probability that a word is *abode*, written $p(abode)$. In the absence of information about context, the maximum likelihood estimate of the probability of a word is just proportional to its frequency. Since *house* is a more frequent word than *abode*, $p(house) > p(abode)$ in the absence of further context, and thus in general *abode* is more informative than *house*. These information-theoretic measures were initially introduced with the description of the linguistic encoding of information as one goal (Shannon, 1948; Bell, 1953; Mandelbrot, 1953; Burton and Licklider, 1955; Pereira, 2000).

However, context can radically alter the informativity of linguistic events. After the sentence fragment *Welcome to my humble __*, the word *abode* is a much more likely completion than *house*. To describe this situation we say the conditional probability $p(abode|humble) > p(house|humble)$. In this specific context, *house* is now *more* informative than *abode*.

Entropy is a measure of uncertainty about an event. It is the weighted average of the self-information of each possible outcome:

$$H(X) = \sum_{x \in X} p(x) log_2 p(x) \tag{1}$$

Suppose a fair coin is flipped, that is $p(\text{heads}) = .5$ and $p(\text{tails}) = .5$. An observer would have maximal uncertainty about the outcome of the coin flip; $H(F) = 1$ bit. But suppose a weighted coin is flipped, where $p(\text{heads}) = 0.9$ and $p(\text{tails}) = 0.1$. In this case there is low uncertainty about the outcome because it is probably heads; calculating entropy, we find $H(F) = 0.49$ bits. It is also possible to calculate the conditional entropy or equivocation, written $H(X|Y)$, which is the uncertainty about an event $X$ given some other event $Y$. For instance, in many languages with grammatical gender, the conditional entropy of gender $G$ given a noun form $N$, $H(G|N)$, is close to zero (Futrell, 2010).

Certain linguistic contexts, however, alter the expected informativity of linguistic material. Certain contexts may leave the language user with more or less conditional entropy about remaining material. For instance, a discourse context such as *I saw the most unusual thing; it was a __* may engender more uncertainty about the noun in the blank than a plain context such as *I saw a* __. More diverse, surprising nouns are likely to follow in the first sentence, inducing greater entropy than the relatively smaller set of nouns that are likely to appear in the second context. This would be reflected in higher entropy for noun phrases in the first context than in the plain context. These effects may arise due to discourse pragmatics, which changes the global context, or due to conventional constructions, which constrain the position of informative material. This thesis has to with the human processing consequences of these information-theoretic properties of English.

### 1.2.2 Processing effects of predictability

The literature on human language processing is unified, as far as I know, on this point: When the words in an utterance are predictable, they are produced and comprehended more easily.

In language production, unpredictable words tend to come along with pauses and disfluencies. The association of informativity with disfluency goes back to Goldman-Eisler (1958, 1961), who estimated predictability using Shannon's (1951) guessing game (played both forwards *and* backwards) and correlated it negatively with the probability and length of pauses in recordings of spontaneous speech. Studies such as Beattie and Butterworth (1979) showed that contextual predictability, rather than general word frequency, predicts these disfluencies most reliably: with predictability held constant, frequency has no effect. Frank and Jaeger (2008) find that the word *the* is lengthened before unpredictable words, likely compensating for the added retrieval time for such words.

The association of disfluencies with informativity is strong enough that comprehenders exploit disfluencies as a cue that upcoming material must be unpredictable. Kidd et al. (2011) show that young children expect previously unmentioned, new referents when they hear disfluencies. In a similar vein to the current research, Corley et al. (2007) find an N400 surprisal effect for *frequent* nouns after disfluencies; they hypothesize that hearers take the disfluency as evidence that a speaker is having trouble retrieving a word, and it would be anomalous for a speaker to have trouble retrieving a frequent word (see also Arnold et al. (2007, 2004)).

For language comprehenders, predictability means faster processing. The finding of faster comprehension for predictable words goes back at least to Marslen-Wilson (1975); Marslen-Wilson and Tyler (1980), who studied the time taken to 'repair' garbled words in various contexts. Eyetracking and reading time studies such as Rayner and Well (1996); Rayner et al. (2001); Staub and Charles Clifton (2006), among many others, have shown that readers are likely to make the effort to fixate on unpredictable words and are likely to skip over predictable ones, even while controlling for length and frequency, two other strong predictors of those effects. The effect of predictability is robust enough that it is standardly controlled for when studying other effects on eye movements (Kliegl et al., 2004).

## 1.3 Predictability affects what is produced

Multiple linguistic forms can often convey the same meaning; in this case the forms can said to be in free variation. In practice, no variants convey exactly the same meaning, so variation is never totally free (Saussure, 1916). However, variant forms are often similar enough in meaning that the decision between them may be influenced by processing factors. Recent work, which I review below, has argued that the choice between these forms may be influenced by information-theoretic properties, with speakers choosing forms that produce a signal with an optimal distribution of information in time.

The recurring theme in this work is that speakers communicate in a way that is easily produced and comprehended, or which enables efficient communication. To do this, they strive to attain something like a constant entropy rate or uniform information density, in which the information conveyed at each time is equal (Levy and Jaeger, 2006). This predicts that uninformative segments should be shortened and informative segments should be lengthened. Furthermore it predicts that additional predictive material should be added to make unpredictable material more predictable in context.

### 1.3.1 Phonetic duration and reduction

One major form of variation in spoken language is in the time and care taken to pronounce words. Jurafsky et al. (2001a,b, 2002) find that word and lemma predictability are among the strongest predictors of duration in spontaneous speech, with more predictable words being pronounced shorter, and with more phonetic reduction. The predictability of a syntactic structure or grammatical construction also affects pronunciation, with more predictable structures yielding a more reduced phonetic signal Gahl and Garnsey (2004); Tily et al. (2009); Kuperman and Bresnan (ress). Predictability is not the only linguistic factor; lemma frequency does seem to lead to shorter pronunciation independently of predictability in some studies (e.g. Gahl (2008)). The reduction and shortening of predictable material, and the lengthening and hyperarticulation of unpredictable material, mean that speakers expend time and effort efficiently.

The precise relationship between informativity and phonetic variation was explored by Aylett and Turk (2004). The authors advance the *Smooth Signal Redundancy Hypothesis*, according to which phonetic reduction and lengthening result in a linguistic signal without moments of excessive or insufficient informativity. That is, the entropy profile of an utterance should be as smooth as possible, without sudden peaks or valleys. In Aylett and Turk's (2004) theory, prosodic prominence provides the mechanism for this smoothing, in that prosodically prominent (in English, stressed) syllables, which are the most informative, are also the ones that expand or contract according to information theoretic pressures. Prominence is then a language's conventional way of dealing with uneven information flow.

### 1.3.2 Using context to increase predictability

Findings about syntactic variation have also been used to support the Smooth Signal Redundancy Hypothesis, albeit under various different names. Word choice and syntactic structure are all influenced by the pressure to avoid over- and underinformativity.

Genzel and Charniak (2002, 2003) support the hypothesis of *entropy rate constancy* by examining the entropy of texts in different parts of discourse. They reason that, at any given time, language users will try to be maximally informative without exceeding some limit on entropy rate. Informativity is unpredictability *in context*, and as a discourse proceeds more and more context becomes available. Therefore, the authors predict that local entropy should *increase* as a text proceeds, because the increase in local informativity is ever offset by increasing available context, which makes the material more predictable globally. For example, an article about cars may take a few sentences to start using phrases like *rotary engine* or *infinite gear ratio*. These phrases are unpredictable and likely overinformative if presented out of the blue. But as the article about cars proceeds, more and more contextual cues become available, so that some sentences later these same phrases might be perfectly predictable. The assumption that predictive context increases and that speakers maintain a maximal entropy rate leads to the hypothesis that linguistic material appears to become more informative as discourse proceeds, but only because it is difficult to measure the rich predictive cues in the context.

The authors find support for this hypothesis in newspaper text. They indeed find that the word-by-word predictability decreases as a function of sentence number in text, with predictability measured according to an n-gram language model. This indicates that writers employ more informative words later in text, once predictive cues have been built up to make those words predictable. They also find that the entropy of syntactic structures increases with sentence number, indicating that not only word choice, but also the *way in which words are arranged* increases in entropy. More recently, Qian and Jaeger (tted) find the same pattern in a sample of 11 different languages.

Building on the hypotheses of entropy rate constancy and smooth signal redundancy, Levy and Jaeger (2006) advance the hypothesis of *Uniform Information Density* (UID), that speakers minimize the peaks and troughs in information rate so as to approximate a uniform density of information in time. The motivating factor is efficiency in communication: a trough in entropy implies wasted effort, while a peak in entropy is potentially incomprehensible to the listener. The authors examine the presence of *that* in contexts such as *The doctor knows that the patient has arrived*; that is, contexts where a verb may be followed either by a nominal object or a sentential complement, where that sentential complement may or not be preceded by *that*. In the face of a variety of controls, the predictability of a sentential complement (as opposed to a direct object) after a given verb is the strongest predictor of whether *that* will appear: if a sentential complement is unlikely, *that* is usually inserted in order to make the upcoming sentential complement more predictable.

In summary, the various smooth-signal hypotheses posit that speakers maximize the effectiveness of their communication by holding the informativity of their message constant over time. This is accomplished by reducing uninformative material, and crucially by using context to make informative material more predictable once it is encountered. Overinformativity is to be avoided because of processing factors: empirically, highly informative material is difficult to process; and theoretically, highly informative material might exceed the capacity of any communication channel, leading to a non-zero probability of errors in transmission (Cover and Thomas, 2006).

## 1.4 Summary

The literature reviewed establishes that predictions about upcoming material are pervasive in language processing, and that the ability to predict upcoming material correctly is a key determinant of processing effort. Furthermore, the predictability-driven processing factors influence language behavior and language structure, preventing elements of language from being excessively informative in context.

In the next subsection I will discuss the limits of predictability as a linear predictor of language processing effort, arguing that we should expect processing difficulty also to arise when linguistic material is *overly* predictable in certain contexts.

# 2 The expectation of informativity

In this section, I lay out the primary hypothesis of this thesis: I argue that speakers have a variable expectation of informativity, in that they exploit linguistic and non-linguistic cues about the informational profile of upcoming discourse. If speakers expect a high information rate at a certain time, but instead are met with highly predictable material, this might itself be an anomaly resulting in processing difficulty.

## 2.1 Problems with predictability

The simple item-by-item predictability of linguistic material may not be the only information-theoretic factor in processing complexity. I investigate this possibility from two angles: the theory of discourse and empirical findings about the distribution of information in English. Theoretically, listeners expect speakers to be informative, which means that listeners must in some cases expect their word-by-word predictions to be frustrated. Empirically, the distribution of information in the speech stream is so uneven that a rational processing system would not assume even predictability, but should rather exploit cues about upcoming informativity and allocate resources accordingly.

### 2.1.1 Expectations about discourse

Predictable material is easy to process, in terms of both production and comprehension. If speakers aim to make their utterances easily processable, or if languages evolve to make crucial information more processable, then they should make linguistic material more predictable. This simple idea leads to a nonsequitur: the most predictable, most easily processed, and thus most ideal utterance would convey no information at all, because information is defined as unpredictability. The sequence *aaaa...* on to infinity is totally predictable, yet cannot convey information. A language whose grammar licensed the production only of the sound [əː] would be similarly optimal. Since the function of language is communication, this conclusion is unpalatable.

Our task in comprehension is not just to predict words, grammatical structures, and meanings, but to figure out the speaker's reasons for saying what he did. For this reason, the monotonic effect of predictability on ease of processing seems problematic in the limit. Intuitively, if someone were to address a listener by saying only [əː] for a few seconds, this utterance, though predictable from second to second, would be extremely difficult to process precisely because the listener would have to do so much pragmatic reasoning to determine the speaker's intentions from very little evidence–work that, for a more sensible utterance, would proceed from far richer cues. An underinformative signal can provoke processing difficulties due to increased reliance on nonlinguistic cues and on pragmatic inference.

Current theories about how processing influences language structure do not predict this kind of extreme reduction. A parallel issue motivates Faithfulness constraints in Optimality Theory. While fulfilling all Markedness constraints would result in all utterances consisting of a single, simple, effortless syllable such as [ba] (Chomsky, 1995), this would violate Faithfulness constraints, which force output forms to be similar to input forms, in effect enforcing the informativity of phonetic forms (Cohen Priva, 2011).

Aside from the prediction of words and meanings in speech, language users may also form more general expectations about what they will hear. They may expect that an utterance will be true, relevant, clear, and informative (Grice, 1975). It must be informative enough to be worth the effort taken to produce it. Zipf (1949) writes: "man talks in order to get something ... words are tools that are used to convey meanings in order to achieve objectives." An uninformative utterance may be anomalous because it is not clear what the speaker is trying to achieve, and it is the inference of the speaker's intentions that drives comprehension. If language comprehenders expect informativity, then overly predictable material might trigger processing difficulties by violating more general communicative expectations.

The expectation of informativity would be well in line with language as it is actually produced. Brown and Dell (1987) find that, in retelling stories involving a given action, subjects only mention the instrument used to perform that action if the instrument is not inferable. If the instrument is the conventional tool for the given action, speakers do not mention it and leave it to the hearer to infer with little effort. A hearer or reader encountering mention of a conventional instrument would thus find the mention of an overly predictable instrument to be out of the ordinary, and requiring pragmatic analysis to explain.

Studies supporting a monotonic effect of predictability often involve language deprived of communicative context. A typical reading time study, such as King and Just (1991), presents subjects with declarative statements such as *The detective that disliked the teacher clipped the coupons in his living room*, without any indication of why a writer might feel the need to inform the reader of such a thing. The communicative anomaly of underinformativity is unlikely to have any effect on reading times with such stimuli, since subjects have no expectation that the sentence was worth producing. In the words of Wittgenstein (1953), we might call this “language on holiday”, and we might expect to find a different, more complicated role for predictability when language is doing more meaningful work in context.

Since underinformativity can be anomalous within an actual discourse setting, predictability might not correlate in a simple way with processing ease, as previous studies have found. In actual discourse, overly predictable utterances might trigger processing difficulties of their own.

### 2.1.2 The informational profile of language

UID states that speakers avoid peaks and troughs in entropy. The reason for avoiding troughs is not that they are difficult to process, but rather that they are a waste of effort. The reason for avoiding peaks is that they may exceed the channel capacity for a given time, resulting in potentially lossy processing. The underlying processing theory is that humans can process a certain constant amount of language information in a given time; in that case it is most efficient for the linguistic signal to approach that constant information rate but not exceed it.

Despite the claim of UID and related theories that language users prefer to approach a smooth entropy rate, the actual entropy profile of language is remarkably uneven. Large spikes in information appear at the beginnings of content words and morphemes, while the later parts of words are almost entirely redundant. A language processor that expects an even and low information rate would not do well in processing such a signal.

The linguistic signal is far from smooth, and an informed processing system could take that into account. If language users form expectations about the informativity of utterances, then they can allocate processing resources at appropriate times and save it at others. Reductions in the amplitude of entropy spikes would still ease processing, but would only be one part of a strategy to keep language efficient and comprehensible.

The first empirical investigation of the informational profile of language was Shannon (1951), who experimentally measured the letter-by-letter entropy of printed English. In Shannon's procedure, the experimenter selects a text and asks a subject to guess what letter will come next in the text. Upon guessing one letter correctly, the subject advances to the next letter. For instance, given $M\ O\ T\ O\ R\ C\ Y\ C\ L$, a subject might guess that the following letter is $E$ in one guess. Given $L\ O\ O\ K\ \_\ A\ T\ \_\ T\ H\ E\ \_$, it might take many guesses for the subject to arrive at the correct next symbol in a given text. The average number of guesses per letter is an upper bound on its self-information or surprisal.

The results of Shannon's guessing game are reprinted in Figure 1 and visualized in Figure 2. Letters requiring one guess convey no information, and 77% of the letters in Shannon's game took 1 guess to determine. For comparison, Cover and King (1978) found an average redundancy of 64% for a different text using their convergent gambling estimate of entropy.

```
T H E R E _ I S _ N O _ R E V E R S E _ O N _ A _ M O T O R C Y C L E _
A _ F R I E N D _ O F _ M I N E _ F O U N D _ T H I S _ O U T _ R A T H
E R _ D R A M A T I C A L L Y _ T H E _ O T H E R _ D A Y
1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1
1 1 3 1 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1 4
1 1 1 1 1 1 11 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1
```

Figure 1: Experimentally derived self-information of English letters in context

Shannon's game is unfortunately played using English letters, which do not straightforwardly represent the actual sounds of English. To determine if the actual speech stream of English contains as much uneven redundancy as the written form, I reproduced the Shannon game using phonetic symbols rather than letters. 3 subjects

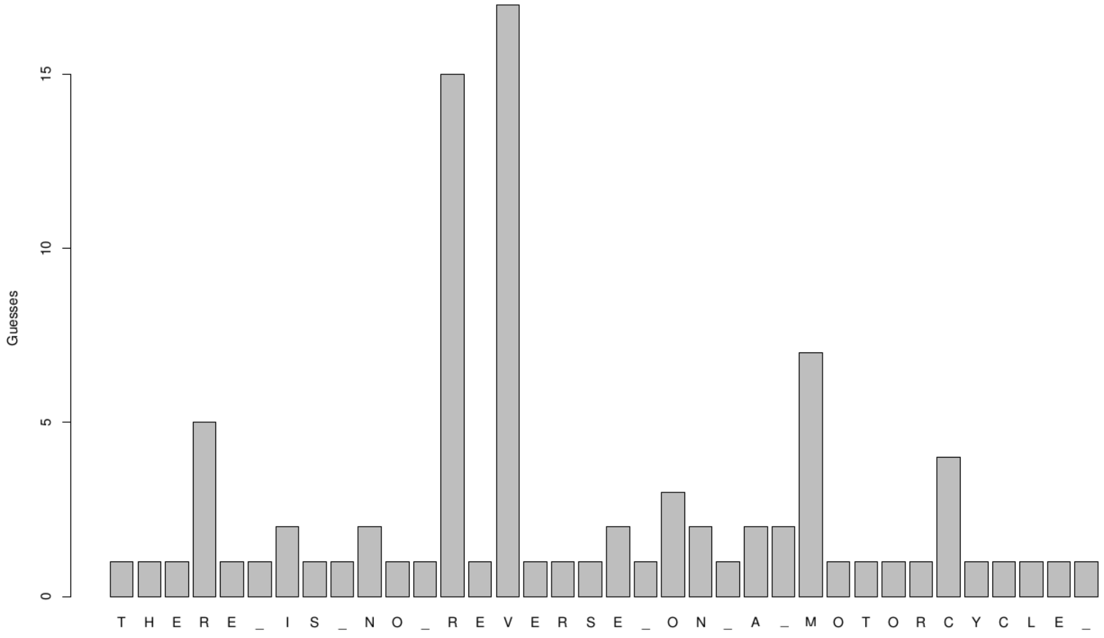

Figure 2: Experimentally derived self-information of English letters in context, visualized.

who were familiar with the International Phonetic Alphabet participated. Part of the results is shown in Figure 3. The large spikes in information at content-word onset, especially at the onset of the first word *reverse*, remain, though the signal does seem smoother than it did with written letters.

Of course, not all phonemes have equal duration. If the redundant phonemes in Figure 3 were a good deal shorter than the informative ones, the signal may yet approach something like a constant entropy rate. However, various empirical findings suggest that the linguistic signal nevertheless consists of swaths of redundancy punctuated with bursts of informativity.

The lengthening and shortening of phonemes is unlikely to eliminate redundancy because phonemes shorten and are dropped not according to their actual predictability in a given context instance, but rather according to their *average* predictability across all their contexts. For instance, while the phoneme /m/ in /əkædəmi/ *academy* is entirely redundant–no other sound can follow /əkædə/ in English–it is very rarely

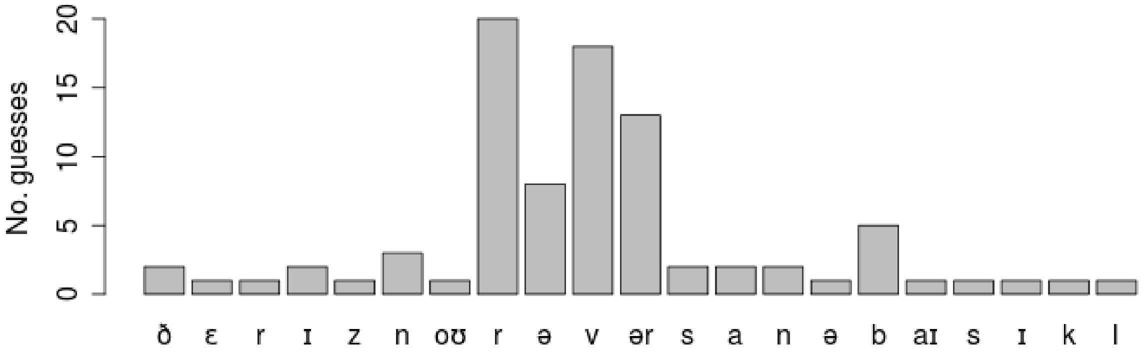

Figure 3: Experimentally derived self-information of English phonemes in context, visualized.

dropped, because in most of its occurrences, /m/ is not redundant (Cohen Priva, 2008).

More directly, the amount of time taken for a listener to identify a word well precedes the end of that word. Marslen-Wilson and Tyler (1980) measured how long subjects took to signal that they had identified words in spoken prose contexts, and found that the response came on average 273 ms after the onset of the word, whereas the average word length was 369 ms. Assuming 50-75 ms are taken to generate the experimental response, the authors conclude that actual word recognition happened about 200 ms into the word. This means that the last 31% of the average word is entirely redundant from the perspective of the language comprehender. [1]

The typical informational profile of the speech stream seems to consist of alternating high informativity at word and morpheme onsets, dropping to zero informativity in the latter third of each word. Production choices influenced by UID might smooth out some of this variance in informational amplitude, but it seems empirically unlikely that they do so well enough to ensure an actually smooth signal. A human

[1] Although word lengths are optimized for efficient information transfer (Piantadosi et al., 2011), this optimization does not apparently eliminate this structured redundancy. Furthermore, the redundancy within words is uneven, with stressed syllables conveying more information (ibid.).

language processing system assuming a constant rate of information transfer would be problematic under these conditions.

Given that the empirical entropy profile of language (or of English, at least) is highly uneven, it would be logical for the language processing system to have mechanisms for the anticipation of bursts of information. The anticipation of informativity is also in line with the pragmatic considerations above: since hearers expect speakers not to waste effort, they must expect a degree of surprise, which means they might in some instances expect their natural word-by-word predictions to be unfulfilled.

## 2.2 Hypothesis: The expectation of informativity

In this thesis I defend the hypothesis that language users are sensitive to cues about the informational topography of linguistic material. That is, language users have a variable expectation of informativity: at some times they expect high information, and at other times they expect low information, allocating processing resources accordingly. This expectation can be modulated by discourse factors and by conventional linguistic structures.

The most relevant previous work on discourse expectations involves information structure, the way in which language users treat new, given, and inferable information. Since new information is unpredictable (otherwise it would be inferable), constructions and contexts that lead language users to expect new material are also leading them to expect a degree of surprise. For instance, Ward and Birner (2004) find that speakers often introduce new discourse referents in frames such as *there was a* __. If the noun in this construction were overly predictable from surrounding context, it might cause more confusion than an unpredictable noun would. At the very least, a predictable noun here would cause more surprise than it would in other contexts, despite appearing in the same discourse.

If speakers have expectations about informativity, then these expectations can be modulated experimentally, and processing difficulties should arise when the actual information profile of an utterance does not match the expected one. The effect should make informative material easier to process, while potentially making uninformative material harder to process. If the effect is particularly strong, then it may even be possible for predictable material to cause *more* processing difficulty than unpredictable material in certain contexts. I report on three Self-Paced Reading (SPR) studies, which measure processing difficulty as word-by-word reading time.

An expectation of informativity could be integrated into the human language processing system in various ways. When language users expect high information, they may simply suppress their expectation of what would be most predictable, or they may simply flatten out their probability distributions over what may follow, making otherwise unpredictable nouns more predictable, and making the most predictable nouns less so. The expectation of informativity may not be mediated through actual word expectations: the processing system may allocate resources in response to an expectation of informativity, rendering everything more easily processed while those resources are available. These implementational details make subtly different empirical predictions and will be discussed in the following subsections.

## 2.3 Empirical predictions

Speakers expect to be surprised because discourse is informative; their expectation of surprise must be modulated in response to various cues because the information content of the linguistic signal is highly uneven. What might happen when this expectation is out of sync with what actually happens? Here I walk through some examples of the hypothesized effects of discourse context and of linguistic constructions whose positioning of old and new information is conventionalized.

One consequence of any form of an expectation of informativity would be that unpredictable words are easier to process in certain contexts, even if those contexts do not directly make the word more predictable. For instance, in a story taking place deep in Amish country, the word *motorcycle* would be highly surprising. It would be quite surprising coming out of the blue, in a sentence such as (1) *As he was walking by his barn, he saw a __*. In this discourse, the word might be less surprising if surrounding context had mutual information with it, for instance if it appeared in the context (2) *An outsider in leather rode down the street in his six-cylinder* __–the context takes some of the edge off of the surprising word *motorcycle.* It also seems that the word might be slightly less anomalous if it appeared in a construction reserving a particular syntactic position for discourse-new material: (3) *As he was walking by his barn, he saw that there was a* __. This last sentence does not make the noun specifically more predictable, as (2) does. Rather, it increases the reader's expectation that an unusual word, refering to an uninferable, new referent, will appear in the blank. If the expectation of informativity plays a role in processing, then *motorcycle* should be easier to process in sentence (3) than in sentence (1).

In a context that induces a high expectation of informativity, informative words should be easier to process and should be read faster than in contexts that induce no such expectation. The effects of *uninformative* words in informativity-favoring contexts is less clear. If the violation of the expectation of informativity itself causes processing difficulties, as does the violation of specific expectations about words, then there should be greater processing difficulty for predictable words in these contexts, as compared to the processing difficulty for those words in other contexts. This effect may even be so large that predictable words become *harder* to process in these contexts than unpredictable words.

The resulting prediction is of an interaction between the predictability of words and the expectation of informativity generated by the word's context. In general, ease of processing correlates positively with a word's predictability. But in a context that favors informativity, unpredictable words become easier to process and predictable words become harder. In an SPR study, with reading time as the dependent variable, we would then predict a negative main effect of predictability, with lower reading times, when there is no expectation of informativity (predictable words have lower RTs). When an informativity-favoring context frame is present, we predict a positive effect of predictability (predictable words have *higher* RTs), or at least a relatively dampened negative effect. This pattern of effects would appear as a *negative interaction* of predictability with context frame, indicating that informativity-favoring contexts reverse, or dampen, the typical effect of predictability.

Further empirical predictions rely on implementational details for the role of expectation of informativity in human language processing. If a context biased a reader to expect informativity, this may lead the reader to allocate processing resources to handle the incoming information. The allocation of resources might mean that all words in expectation-inducing contexts would be read faster. This would manifest as a negative main effect of context, with faster reading times for both predictable and unpredictable words.

Processing effects of the expectation of informativity might be mediated through comprehenders' specific expectations about upcoming material. For instance, in a context where a comprehender expects new material, her probability distribution over words that may simply be flattened out, so that all words are relatively equally predictable. In that case, the relatively predictable words would be less so, and the relatively unpredictable words would be more predictable than in a context that does not specifically favor new material. This would produce the same negative interaction that we predicted from the violation of the expectation of informativity, except that this negative interaction could not result in a crossover where predictable words are harder than unpredictable ones. If the crossover is found, then we could conclude that this implementation is not the only way that language users attend to cues about information.

The primary prediction is a negative interaction between predictability and context, whereby informativity-biasing contexts reverse the effects of predictability. Below, we report on three self-paced reading studies, aiming to test for this novel effect using both discourse and conventional contexts, and to determine how the reading time data can inform us about the specific role of cues about informativity in human language processing.

# 3 Using Mechanical Turk for Self-Paced Reading studies

The reading time stimuli we developed were highly formulaic, and it is likely that an experimental subject might figure out very quickly what was being manipulated. For this reason, we decided to run the self-paced reading study over Mechanical Turk, which allowed us to gather 136 unique subjects per study in a reasonable amount of time; this way each subject would only be exposed to a few critical stimuli.

I am not aware of any published work using Mechanical Turk for the collection of self-paced reading time (SPR) data. For that reason, I will describe my procedures for using Mechanical Turk in detail here, as well as presenting results from two validation studies replicating classic psycholinguistic results over Mechanical Turk.

## 3.1 About Mechanical Turk

Amazon Mechanical Turk is a 'crowdsourcing' platform in which workers ("Turkers") perform small tasks as contractors, for small amounts of money. These tasks are called "Human Intelligence Tasks" (HITs), and typically involve some process that computers cannot yet do reliably, such as image tagging. Each HIT consists of multiple assignments, and a given subject may not complete more than one assignment within a HIT. A Turker accepts an assignment, completes work, then submits the results to Amazon's servers, whereupon the work requester can choose to approve the work, thus paying the worker, or to reject it.

The large number of workers on Mechanical Turk means that hundreds of assignments can be completed in days, at relatively low cost (the average Turker makes about $1.00 per hour from HITs).

The population of Turkers provides a varied if not balanced sample of the population of native American English speakers. Most Turkers are either American (46.8%) or Indian (34%), since the United States and India are the only countries where Turkers can be paid in cash rather than in Amazon.com gift cards. Among Americans, most Turkers are females seeking secondary income, spending on average 4-8 hours per week on Mechanical Turk for an average weekly income of about $7. Ages cover a large range, though most workers are between the ages 20-40. 80% of Turkers have at least a Bachelor's degree (Ipeirotis, 2010). The distribution of ages in particular makes the Turk population more representative of the general populace than a college undergraduate population.

## 3.2 Previous linguistic and behavioral research using Mechanical Turk

The population of Mechanical Turk workers has proven reliable for linguistic annotation tasks and for behavioral experiments. Overall, when proper measures are taken to filter out unreliable or deceptive workers, the quality of results is just as good as that from controlled academic settings.

In the language domain, crowdsourced data has been evaluated in NLP studies such as Snow et al. (2009), which used Turkers to annotate texts for affect, word similarity, textual entailment, temporal ordering, and word-sense disambiguation. This study found high levels of agreement between academic and crowdsourced annotators, with accuracies as high as 99.95% for the word-sense disambiguation task. The crowdsourced labels were often as reliable as gold-standard annotations for supervised machine learning tasks. Hsueh et al. (2009) and Callison-Burch (2009) also provide useful evaluation of annotation and translation quality from Mechanical Turk.

Reports on linguistic and psycholinguistic studies using Turkers as a population also find that the quality of work is entirely satisfactory. Munro et al. (2010) report on several studies, including the one above. For word segmentation, cloze-task predictability estimation, and grammaticality judgments, results from Mechanical Turk mirrored those conducted in academic labs. Sprouse (2011) provides another validation of the use of Mechanical Turk to collect grammaticality judgments. When asked to rate the semantic transparency of phrasal verbs, Turkers have an interrater reliability of $K = 0.823$, or ‘almost perfect agreement’ according to the criteria in Landis and Koch (1977). The intra-class correlation (ICC), a measure of agreement among participants, was very high at 0.78% in the case where each individual Turker rated

18 verbs; the ICC dropped to 0.09 when Turkers were asked to rate 96 (Schnoebelen and Kuperman, 2010). This result suggests that it is wise to recruit Turkers for only a few tasks at a time.

The collection of data for on-line processing is somewhat different from the studies reported above, because it requires the collection of reaction time data. This introduces some potential problems for crowdsourced data: different internet speeds and computer hardware might result in inconsistencies in reaction time. For instance, if a subject has a slow internet connection and a slow computer, she may see the words of the self-paced reading display more slowly, resulting in artificially slow reported reaction times. It also remains unclear whether a web-based reaction time collection system can accurately measure small time intervals.

Keller et al. (2009) report on a number of experiments to address these issues for the web-based distribution of reaction time experiments, though they do not use Mechanical Turk. Using their WebExp system, they found time measurements to be mostly constant regardless of platform and load. For instance, the experimenters conducted an experiment where a key was pressed every 300ms on a computer simultaneously running Yahoo! Messenger, MSN Messenger, Zone Alarm, Norton Antivirus, and Google Video; the web-based software measured the time between presses within an accuracy of 20ms. The authors also replicated Sturt et al. (1999), a self-paced reading time study, finding the same pattern of results as the original paper and an overall correlation of $r = 0.455$ with the original reading times. With a low rate of visitors to the website hosting the experiment, this replication took 8 months to complete.

These results indicate that web-based experiments are a reliable way to conduct self-paced reading time experiments, but it remains to be seen whether a web-based experiment distributed to Mechanical Turk workers is reliable. Below, I describe my method for distributing experiments, as well as two validation studies replicating well-known findings.

## 3.3 How to run SPR experiments on Mechanical Turk

In this subsection I describe my procedure for distributing self-paced reading time experiments over Mechanical Turk. I have also created a detailed tutorial on how to set up experiments, accessible on my website at <http://stanford.edu/ rfutrell/mt.html>.

The actual software used to implement the experiment is Alex Drummond's *Ibex* [2], formerly known as *webspr*. The software allows for both dashed-window and centered display of words. These web experiments can be hosted at *Ibex Farm* [3] free of charge. With the experiment hosted and operational, the only remaining task is to link it to Mechanical Turk so that workers can see the experiment and submit the completed assignment to the Mechanical Turk servers upon completion.

Typically, when running HITs on remote servers, work requestors will ask Turkers to navigate to an external site, complete a task, then receive a password which they reinsert into the Mechanical Turk website to verify that they have completed the task. I chose to use a simpler method. The Mechanical Turk Command-Line Tools (CLT) provide tools for running 'external HITs', where data from an external website appears within the Mechanical Turk interface. For the Turker, it does not appear that he has left the Mechanical Turk website at all. When the experiment is completed, the Turker submits the HIT as usual without the need to copy a password or other means of verification. For about 2% of workers, this external HIT fails to operate as planned, and they are unable to submit the assignment for money; I compensated these workers for their time by paying them a bonus.

## 3.4 Validation studies

I conducted two validation studies to ascertain whether a web-based SPR experiment distributed over Mechanical Turk could detect well-known psycholinguistic phenomena. The first study is a replication of frequency effects with stimuli from Rayner and Duffy (1986); the second is a replication of the asymmetry between subject relative and object clauses with simuli from King and Just (1991).

[2] http://code.google.com/p/webspr/

[3] http://spellout.net/ibexfarm/

### 3.4.1 Frequency effects

The finding that infrequent words are read more slowly than frequent words is one of the most robust in psycholinguistics. If an SPR experiment over Mechanical Turk failed to detect this effect, it would be extremely damning for crowdsourcing as a way to gather SPR data.

**3.4.1.1 Materials** I selected stimuli from a typical psycholinguistic paper exploring frequency effects, Rayner and Duffy (1986), which gathered eye-tracking data. In their Experiment 1, the authors study three effects: noun frequency (*chicken* versus *rooster*), verb semantic complexity (*kill* versus *die*), and local syntactic ambiguity. They find slow reading times for infrequent nouns and for syntactic ambiguities but not for semantically complex verbs. This was the first study to study word frequency while controlling for word length: all the authors' infrequent nouns are just as many letters as their frequent nouns.

All stimulus sentences are presented in Figure 4.

The old (gondola/vehicle) creaked loudly.
The large (mosque/church) remained mostly empty.
The plump (rooster/chicken) chased the sparrows.
The beautiful (dunes/beach) stretched for many miles.
The young (waiter/driver) annoyed his friends.
The lovely (waltz/music) seemed out of place.
The angry (refugee/officer) ignored the food.
The tall (steward/student) left the plane.

Figure 4: Stimulus sentences from Rayner and Duffy (1986).

**3.4.1.2 Method** I used Rayner and Duffy (1986)'s noun frequency stimulus sentences in a self-paced reading time experiment in Ibex, distributed over Mechanical Turk. The format was a word-by-word dashed moving window display, in which subjects see a sentence profiled in dashes with one word revealed in text. Upon clicking the mouse or pressing any key, that word is obscured in dashes and the next word is revealed, while reaction times between clicks or key presses are recorded on the Ibex server.

80 unique subjects were recruited for the experiments. In order to keep the study short and thus attractive to Mechanical Turk workers, each subject only saw 4 of the 8 critical sentences, with each subject seeing two low-frequency (lf) words and two high-frequency (hf) words. Each subject also saw 5 filler sentences for a total of 9 sentences per subject. Each sentence was followed by a comprehension question to ensure attentiveness.

Workers were paid $0.60 per assignment, a relatively high rate on Mechanical Turk. Demographic data was collected for each subject: gender, age, and native language ("What language(s) did you grow up speaking?"). All subjects were required to indicate their consent to the experimental protocol before proceeding.

Special care was taken to remove suspect subjects and trials. All subjects who entered a native language other than solely "English" were excluded from the data analyzed; nevertheless, all workers who completed the HIT were paid. Subjects with a mean RT four SD above or below the overall mean, or with a mean RT below 100 ms, were excluded after inspection to ensure that the abnormal mean RT was not due to a few outlying RTs. Subjects scoring less than 1 SD below the mean accuracy on comprehension questions were excluded. Any trial containing multiple contiguous reaction times under 50 ms (indicating that the subject was simply holding down a key) was excluded.

**3.4.1.3 Results** Data collection took 4 days. 10 subjects were excluded for not listing "English" as their native language. A histogram of mean RTs for subjects is shown in Figure 5; no subjects were removed for high or low mean RTs. Comprehension question accuracy was not used as a filter for this experiment; because of the short nature of the stimulus sentences, almost all subjects answered every question correctly, and no subject got more than one question wrong.

The resulting reading times show a clear effect of frequency. Figure 6 shows the raw mean reading times for each word in each condition (red or lf is low-frequency, while black or hf is high-frequency). Figure 7 shows the results after residualizing on

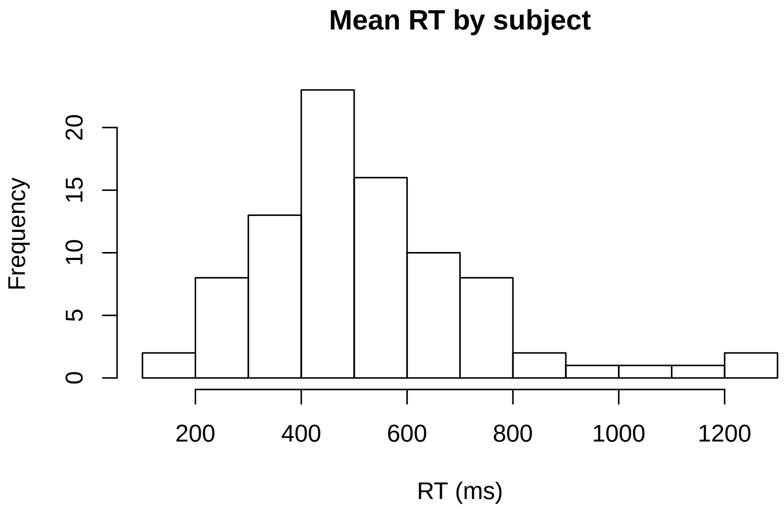

Figure 5: Histogram of mean RTs by subject

word length.

The low-frequency nouns were read on average 121 ms slower than the high frequency nouns. Rayner and Duffy (1986) found that subjects fixated on the low-frequency nouns for about 40 ms longer than the high-frequency nouns; however the effect sizes are not immediately comparable since the previous study was using eye-tracking while this study is using SPR. Table 1 shows the results of a mixed-effects model of the data, with subject and item as random intercepts, predicting reading time of the critical word from its frequency class (low or high). The effect of frequency is significant at $p = 0.023$.

| Coefficients: | Est. Std. | Error | $t$ value | Pr(>$\|t\|$) |
|---|---|---|---|---|
| (Intercept) | 538.6 | 47.5 | 11.3 | <0.001 |
| low frequency | 121.7 | 53.2 | 2.3 | 0.023 |

Table 1: Mixed-effects linear regression predicting the RT of the critical word. $p$ values obtained by Markov Chain Monte Carlo bootstrapping.

The data indicate that a web-based SPR study over Mechanical Turk can replicate basic frequency effects. With this basic sanity check out of the way, we can move

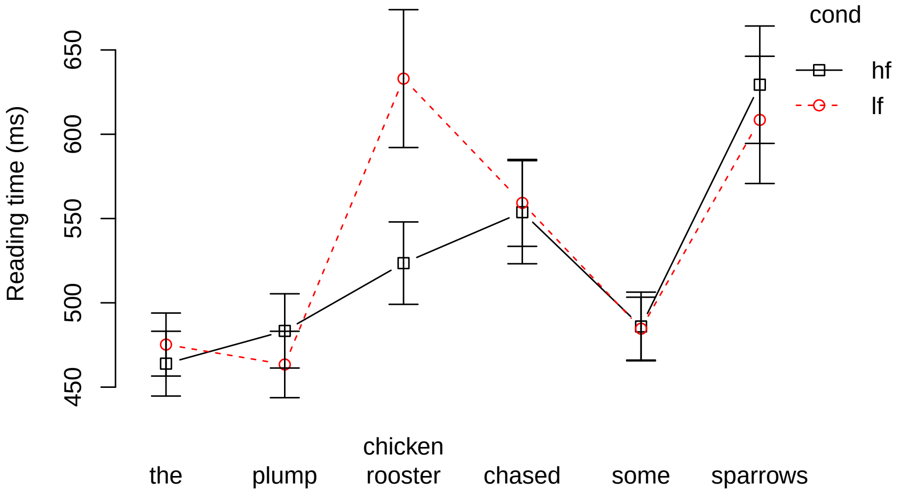

Figure 6: Raw mean RTs by word

on to a second validation study of another well-known psycholinguistic result: that object-relative clauses are harder to process than subject-relative clauses.

### 3.4.2 Subject and object relative clauses

The processing asymmetry between subject and object relative clauses is a well-known finding with a variety of different explanations (Gibson, 1998). Empirically, the phenomenon is slower reading times for both the matrix verb and the embedded verb in sentences with object-relative clauses, such as *The reporter that the senator attacked admitted the error*, as compared to reading times for the same verbs in subject-relative clauses, such as *The reporter that attacked the senator admitted the error*. Since this is such a robust and well-known effect, it provides another useful validation for the use of web-based SPR over Mechanical Turk.

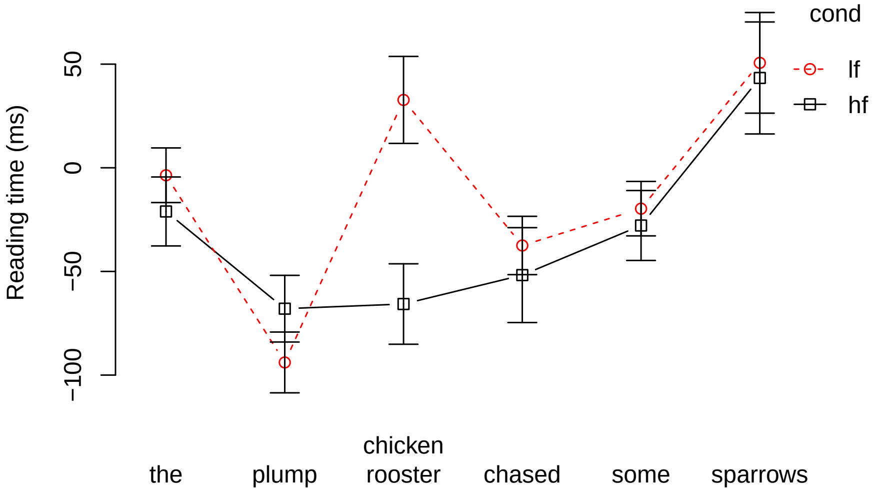

Figure 7: Residual mean RTs by word, controlling for word length

**3.4.2.1 Materials** Sentences with subject- and object-relative clauses were taken from King and Just (1991)'s Experiment 2. The original study investigated the effects of working memory span on reading time and comprehension accuracy for sentences with these relative clauses. The study also investigated the role of mutual information between verbs and their subjects/objects by using verbs biased to take certain nouns as subject or object; i.e. *instructed* is more likely to take *teacher* as its subject than *robber*. I used the 8 stimulus sentences from the original materials, in which neither the embedded nor matrix verbs had a strong bias towards the various nouns in the sentence. The stimuli are presented in Figure 8.

**3.4.2.2 Methods** I used King and Just (1991)'s stimulus sentences in a self-paced reading time experiment in Ibex, distributed over Mechanical Turk. The format was a word-by-word dashed moving window display.

100 unique subjects were recruited for the experiment. Each subject only saw 4 of

The banker that (praised the barber/the barber praised) climbed the mountain just outside of town.
The lawyer that (phoned the reporter/the reporter phoned) cooked the pork chops in their own juices.
The salesman that (liked the fireman/the fireman liked) dominated the conversation about the pennant race.
The waiter that (despised the broker/the broker despised) drove the sportscar home from work that evening.
The detective that (disliked the teacher/the teacher disliked) clipped the coupons out with the dull scissors.
The judge that (ignored the doctor/the doctor ignored) watched the movie about Colombian drug dealers.
The robber that (insulted the accountant/the accountant insulted) read the newspaper article about the fire.
The governor that (admired the comedian/the comedian admired) answered the telephone in the fancy restaurant.

Figure 8: Stimulus sentences from King and Just (1991).

the 8 critical sentences, with each subject seeing two subject-relative sentences (*srel*) and two object-relative sentences (*orel*). Each subject also saw 5 filler sentences for a total of 9 sentences per subject. Each sentence was followed by a comprehension question.

Workers were paid $0.70 per assignment. Demographic data was collected for each subject: gender, age, and native language ("What language(s) did you grow up speaking?"). All subjects were required to indicate their consent to the experimental protocol before proceeding.

All subjects who entered a native language other than "English" were excluded from the data analyzed; nevertheless, all workers who completed the HIT were paid. Subjects with a mean RT four SD above or below the overall mean, or with a mean RT below 100 ms, were excluded after inspection to ensure that the abnormal mean RT was not due to a few outlying RTs. Subjects scoring less than 1 SD below the mean accuracy on comprehension questions were excluded. Any trial containing multiple contiguous reaction times under 50 ms (indicating that the subject was simply holding down a key) was excluded.

**3.4.2.3 Results** Data collection took 3 days. 4 subjects were excluded due to not entering "English" as their native language. Data for one subject-relative sentence was excluded due to a typo. 16 subjects were removed due to poor scores on comprehension questions (below 71%). 2 subjects were removed due to abnormally high mean RTs. The relationship between mean RT for subjects and comprehension scores is shown in Figure 9: subjects with faster RTs generally did poorer on comprehension questions, indicating that some subjects were not as attentive as could be desired.

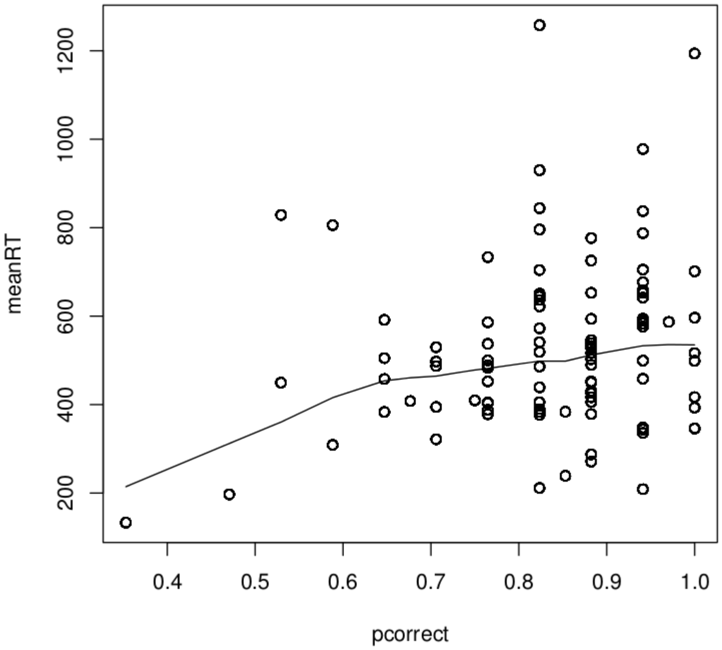

Figure 9: Mean RT by proportion of correct answers for subjects. A locally-weighted least squares regression line is fitted.

The RT results showed a significant slowdown at the embedded and matrix verbs of sentences with object-relative clauses. Figure 10 shows the raw mean reading times for each word in each condition (red or srel is subject-relative, while black or orel is object-relative). Figure 11 shows the results after residualizing on word length.

The embedded verb in a subject-relative clause was read on average 209 ms faster than the same verb in an object-relative clause; for the matrix verb the average difference is 202 ms. This is in keeping with King and Just (1991), who find an effect

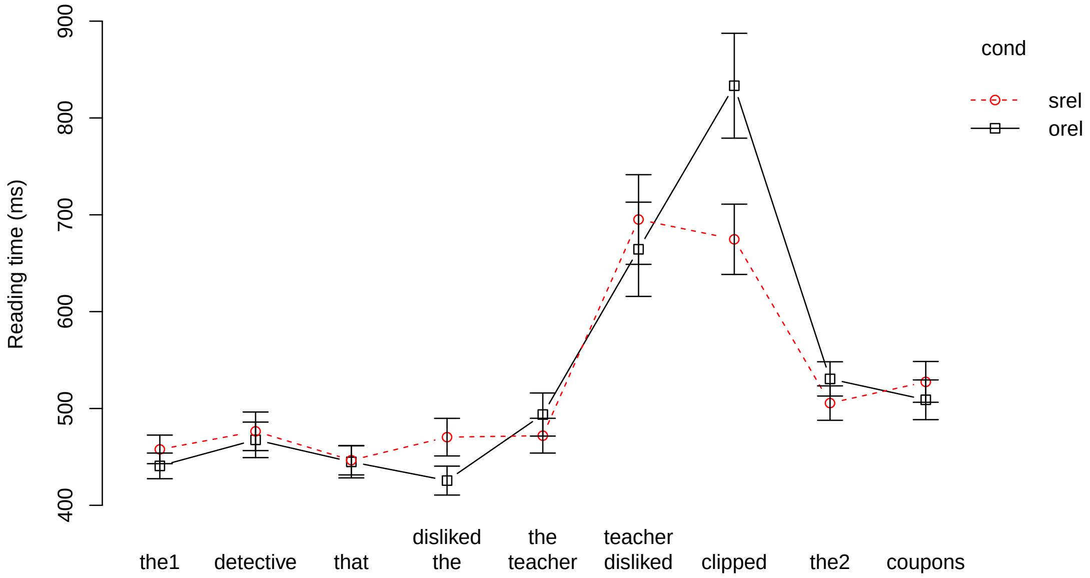

Figure 10: Raw mean RTs by word

size of about 200 ms for the matrix verb; the effect size for the embedded verb here is larger than King and Just (1991)'s finding of about 100 ms difference. Mixed-effect models, summarized in Tables 2 and 3 with subject and word as random intercepts confirm the significance of both of these effects. Overall, the results comprise a satisfactory replication of this finding.

| Coefficients: | Est. Std. | Error | $t$ value | Pr(>\|$t$\|) |
|---|---|---|---|---|
| (Intercept) | 910.7 | 68.0 | 13.4 | <0.001 |
| subject-relative | -202.6 | 59.6 | -3.4 | <0.001 |

Table 2: Mixed-effects linear regression predicting the RT of the matrix verb. The intercept represents the RT for object-relatives. $p$ values obtained by Markov Chain Monte Carlo bootstrapping.

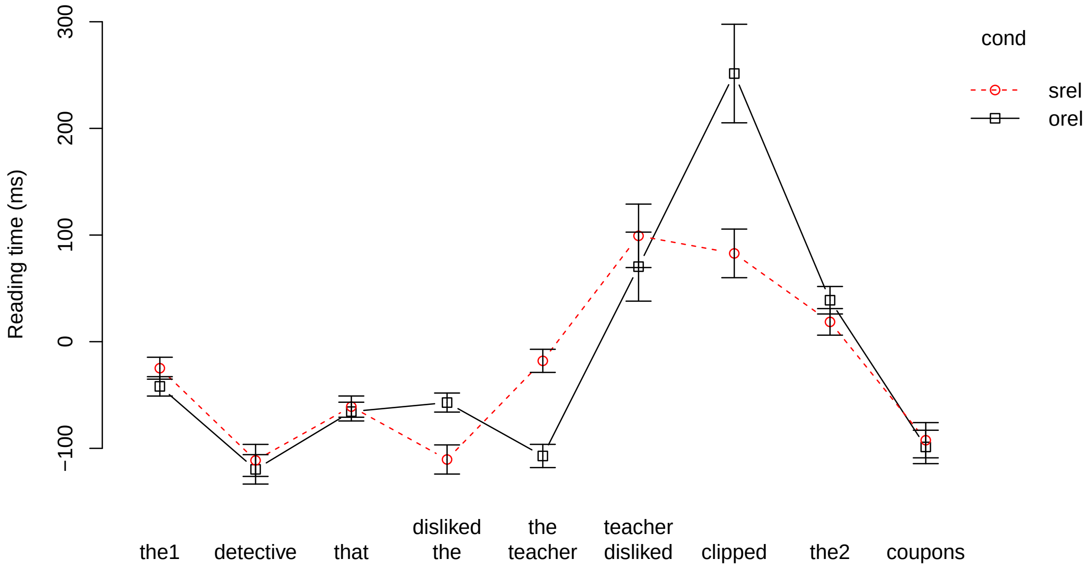

Figure 11: Residual mean RTs by word, controlling for word length

| Coefficients: | Est. Std. | Error | $t$ value | Pr(>\|$t$\|) |
|---|---|---|---|---|
| (Intercept) | 910.7 | 68.0 | 13.4 | <0.001 |
| subject-relative | -202.6 | 59.6 | -3.4 | <0.001 |

Table 3: Mixed-effects linear regression predicting the RT of the matrix verb. The intercept represents the RT for object-relatives. $p$ values obtained by Markov Chain Monte Carlo bootstrapping.

**3.4.2.4 Summary** Mechanical Turk allows a large number of subjects to be recruited for an experiment at low cost and in short time. This allows us to proceed in this thesis research using stimuli that repeatedly violate discourse expectations without worrying that readers become accustomed to such violations; there are enough subjects that each subject may see only a few critical stimuli. More generally, I hope the literature review and validation studies of this subsection enable people to use Mechanical Turk for future SPR studies with confidence. The quality of data seems to be entirely satisfactory for behavioral studies.

# 4 Experiments

Three self-paced reading (SPR) experiments were carried out to determine the effects of the expectation of informativity, and of the violation of that expectation. The first and second SPR experiments deal with the effect of discourse context, in which knowledge about the topic of discussion influences a reader's expectation about what might be surprising. The third SPR experiment deals with the effect of a conventional focusing construction, the wh-cleft, which we hypothesize promotes an expectation of informativity.

## 4.1 Description of stimuli

All our SPR stimuli involve a conflict between global and local expectations. In this sense, the setup is similar to Nieuwland and van Berkum (2006), in which global context establishes a peanut as an animate discourse referent, overriding hearers' prior expectations about the use of the inanimate noun *peanut*. In their experiment, hearers' expectations are shifted such that they expect animate descriptions of the peanut, rather than inanimate descriptions. That is, hearers expect a certain set of words other than the set of words they would have expected without context. In our case, global context suggests to readers that a *surprising* word will be used. This is different than shifting one's expectations from one set of words to another; in this case one's expectations must be shifted from a small set of predictable words to a potentially unbounded set of unpredictable words.

We set up stimuli in which a subject is described using an instrument that is either conventional or unconventional for a task. We chose to use instruments as the loci of surprise based on previous work on conventional instruments such as Brown and Dell (1987), who demonstrated that speakers avoid mentioning instruments if they are underinformative.

A typical stimulus is structured as in Figure 12, in which the reader is led to expect surprise then encounters a surprising, unconventional instrument. A discourse is established about *John* and his surprising ways of accomplishing tasks, using the phrase *John is a surprising person who never does things the way you'd expect.* This establishes an expectation that following material will contain some surprising information about the way John does something. A following sentence then mentions some task that has a well-defined conventional instrument, and claims John uses an unconventional, surprising one: *For instance, in order to brush his teeth, he was using a knife the day before yesterday.* In the figure, green checks represent fulfilled expectations and red exclamation points denote unpredicted events. The unpredictability of the unconventional instrument word *knife* itself fulfills the discourse expectation. The discourse expectation does not make *knife* directly more predictable: all it does is establish that *some* unexpected instrument will be used. The stimulus is then finished with a four-word time adverbial such as *the day before yesterday* (not shown in the figure) to provide spillover words for reading time measurement.

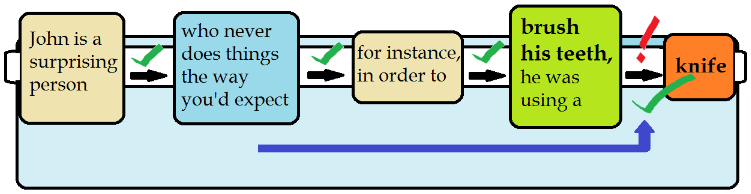
John is a surprising person
who never does things the way you'd expect
for instance, in order to
brush his teeth,
he was using a
knife

Figure 12: Structure of a typical stimulus with congruent discourse expectations. A carrier phrase induces an expectation of surprise. The unconventionality of the instrument word validates that expectation.

A stimulus from another condition is depicted in Figure 13, where readers are led to expect surprise, but instead encounter only a conventional instrument. The discourse again begins by establishing that John and his surprising methods are the topic under discussion. This time, however, the action is changed, so that the instrument is conventional and thus entirely predictable: *For instance, in order to chop some carrots, he was using a knife the day before yesterday.* The *predictability* of the word *knife* violates the discourse expectation of surprise.

Other stimuli established the opposite kind of expectation, inducing readers *not* to expect especially informative discourse. In this case, we would expect the classic slowdown effects for unpredictable instruments. Examples of stimuli from the four total

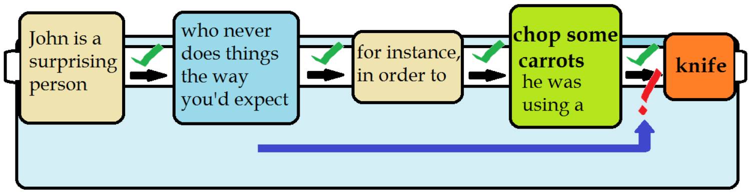
John is a surprising person
who never does things the way you'd expect
for instance, in order to
chop some carrots he was using a
knife

Figure 13: Structure of a typical stimulus with incongruent discourse expectations. A carrier phrase induces an expectation of surprise. The conventionality of the instrument word violates that expectation.

conditions are given in Figure 14. We prepared a set of 13 action–conventional instrument pairs, and shuffled the instruments to create a set of 13 action–unconventional instrument pairs. Thus all instrument words appeared for one typical action and one atypical action. The resulting 26 action-instrument pairs are shown in Table 4. These 26 resulting sentences appeared in two kinds of discourse context: one that promoted an expectation of informativity (*surprise*) and one that promoted no such expectation (*boring*). Examples of all four resulting conditions are shown in Figure 14.

| Action | Conventional instr. | Unconventional instr. |
|---|---|---|
| eat steak | fork | pen |
| dig hole | shovel | fork |
| chop carrots | knife | shovel |
| brush teeth | toothbrush | knife |
| clean porch | broom | toothbrush |
| repair brakes | wrench | broom |
| secure yacht | rope | wrench |
| accessorize dress | belt | rope |
| transport groceries | cart | belt |
| drain spaghetti | strainer | cart |
| wrap present | ribbon | strainer |
| wash dishes | sponge | ribbon |
| write letter | pen | sponge |

Table 4: Action-instrument pairs for the conventional and unconventional conditions.

Because the stimuli are highly formulaic–only the action and instrument words are

**Expect surprise, get unconventional:** John is a surprising person who never does things the way you'd expect. For instance, in order to brush his teeth, he was using a knife the day before yesterday.

**Expect surprise, but get conventional:** John is a surprising person who never does things the way you'd expect. For instance, in order to chop some carrots, he was using a knife the day before yesterday.

**Expect boring, get conventional:** John is a boring person who always does things the way you'd expect. For instance, in order to chop some carrots, he was using a knife the day before yesterday.

**Expect boring, but get unconventional:** John is a boring person who always does things the way you'd expect. For instance, in order to brush his teeth, he was using a knife the day before yesterday.

Figure 14: Stimulus conditions and examples for SPR studies.

substituted–subjects could only be exposed to a few critical stimuli, lest they become inured to the manipulation. Running the experiments over Mechanical Turk allowed us to gather data from a large number of unique subjects who could be recruited efficiently for short experiments. Each subject only saw four critical stimuli, one from each condition.

The construction of the stimuli aims to discover the effects of a relatively realistic discourse context that leads readers to expect informativity. We seek especially to induce violations of that expectation, to determine what kind of effects, if any, may follow. From the discussion above, we predict that the expectation of informativity should be detrimental to predictable, conventional nouns (increasing their reading time) and beneficial for surprising, unconventional nouns (decreasing their reading time).

One assumption in the construction of our stimuli is that the unconventional instrument nouns are unpredictable given the action described. If one expects John to be brushing his teeth with some unconventional instrument, then we assume one does not have specific expectations about what the instrument will be. Rather, one must believe a fairly uniform distribution of instruments may follow. Thus any following instrument will be relatively unpredictable.

This assumption may not be wise, however, in that readers may form specific expectations about unconventional instruments. When expecting an unconventional instrument for the action *stirring coffee*, for instance, they may expect *finger* or *twig*; that is, they may expect a set of 'conventionally unconventional' instruments for certain actions. The following section describes a sentence completion study we conducted in order to confirm the assumption that the expectation of surprise leads to an expectation of a genuinely surprising set of instruments.

## 4.2 Experiment 1: Sentence Completion

Since our studies aim to discover the interaction of actual unpredictability with discourse expectations, we would prefer that the unconventional instruments in our stimuli be actually unexpected. For this reason, before constructing stimuli for SPR, we carried out a sentence-completion study to determine what instruments were 'conventionally unconventional' for what actions. The sentence completion task also helped us ascertain the most predictable instruments to be mentioned for certain actions. Since speakers do not usually say *brushing teeth with a toothbrush*, it is possible that the instrumental PP actually does *not* make *toothbrush* predictable in that context, if hearers' expectations are based solely on frequency in experience.

We expected that adding discourse context would lead to a more informative distribution of completions, in that more unconventional instruments would be selected and that more varied instruments would be selected.

### 4.2.1 Materials & Methods

23 incomplete sentences were presented either with or without discourse context to Amazon Mechanical Turk (AMT) workers, who were asked to complete them. A sample fragment is *Aaron was opening a bottle with* __, which was either presented alone or preceded by the sentence *You'll never believe it!*, providing discourse context and an expectation of informativity. The 23 fragments appeared either with or without discourse context for a total of 46 individual stimulus types. AMT workers were paid $0.05 per completion; no worker completed more than 5 separate sentences. AMT workers were asked for their native language and results were thrown out if this

was not English. Response times two standard deviations greater than the average response time were also removed from the dataset. After this filtering, 786 valid completions remained, with an average response time of 22.9 seconds.

### 4.2.2 Results

The AMT results revealed that conventional instruments were the most likely continuations of sentences without discourse context, while a wide set of unconventional instruments were likely as continuations of sentences with discourse context. For the condition without discourse context, 349/399 (87.5%) of completions were conventional; with discourse context, only 161/387 completions were conventional (41.6%; $\chi^2$ = 38.8, $p$ <0.001).

Discourse contexts led writers to produce more informative completions, validating the possibility of an expectation of informativity in this context. The average entropy of the distribution of instruments after discourse context was 3.5 bits. Without discourse context, the average entropy was 2.8 bits. The difference between mean entropies was significant ($t(37.0)$ = -4.1, $p$ <0.001). The difference in entropies means that there is greater uncertainty about exactly what instrument will follow when there is discourse context, and less uncertainty when there is no discourse context.

The average entropy of unconventional instrument completions was 3.1 bits; for conventional completions the mean entropy was 1.8 bits ($t(32.6)$ = -5.3, $p$ <0.001). Figure 15 shows the entropies for sentence completions in different conditions: for almost all sentence prefixes, the entropy in the discourse-context condition is higher than the entropy without discourse context.

A set of 'conventionally unconventional' instruments sometimes appeared, resulting in a relatively low informativity for unconventional completions. For instance, the sentence prefix *Josh was eating spaghetti with* had an entropy of only 2.3 bits with discourse context, compared to 2.0 bits without discourse context. Without discourse context, subjects completed with sentence overwhelmingly with *a fork*. With discourse context, subjects completed the sentence mostly with *a knife* and some with *a spoon*. When subjects expect an unconventional instrument in this sentence, they might have very little uncertainty about what that instrument will be, because it is

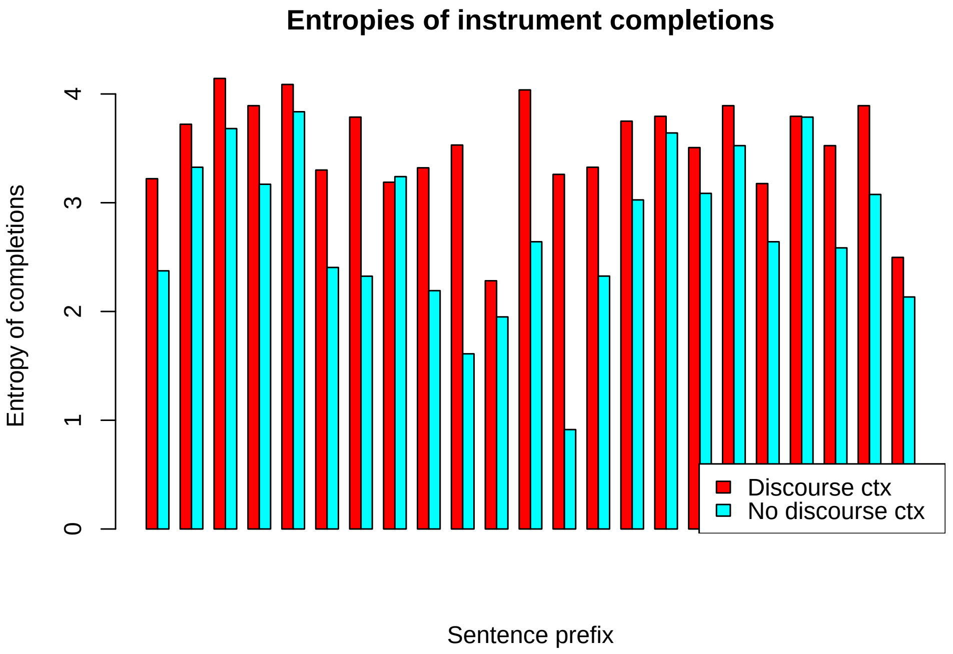

Figure 15: The entropy of instrument completions for each sentence prefix, with and without discourse context.

apparently almost always *knife*. 'Conventionally unconventional' instruments often involved body parts, such as brushing teeth or stirring coffee with a finger. If any one response dominated the unconventional completions for a sentence, that response was noted so it could be avoided in future stimulus construction.

The sentence completion task demonstrated to our satisfaction that certain discourse contexts can lead to an expectation of informativity, and allowed us to determine which actions had the starkest difference in informativity between conventional and unconventional completions.

## 4.3 Experiment 2

In this experiment, we attempted to induce an expectation of informativity using discourse context, then to violate that expectation. The results from this first experiment are confusing, but they become explicable when considering results from subsequent experiments.

In order to create discourse expectations, stimuli were constructed either with the leading sentence *You'll never believe this!* in the "expect surprise" condition, or with the sentence *What a boring week* in the "expect boring" condition.

The challenge is then: how to *localize* readers' expectation of informativity? That is, how to cue to readers exactly *where* the informative material should be, so that we can observe surprise when the material there isn't informative? If readers expect that something about the following sentence will be informative, then they will be unfazed by uninformative material, because they might expect that they simply haven't reached the informative material yet. The expectation of informativity can only be violated if readers have specific expectations about what material will be surprising; without that expectation, readers will simply keep expecting that the surprising material is still upcoming.

We attempted to localize the expectation of informativity using a cleft construction. In the condition where speakers expect surprise and get it, and where speakers don't expect surprise and don't get it, a stimulus is of the form:

**Expect surprise, get unconventional:** You'll never believe it! The thing John was brushing his teeth with was a knife the day before yesterday.

**Expect surprise, but get conventional:** You'll never believe it! The thing John was chopping some carrots with was a knife the day before yesterday.

**Expect boring, but get unconventional:** What a boring week. The thing John was brushing his teeth with was a knife the day before yesterday.

**Expect boring, get conventional:** What a boring week. The thing John was chopping some carrots with was a knife the day before yesterday.

The aim was to lead readers to believe that informative material would appear after *the thing [NAME] was [ACTION] with was a* .

### 4.3.1 Materials

The 13 actions in Table 4 appeared with both conventional and unconventional instruments, in discourse contexts leading either (1) to an expectation of surprise (the *expect surprise* condition) or no expectation of surprise (*expect boring*), for a total of 13 x 4 = 52 stimuli.

We expected the fastest reading times when readers expect surprise and get an unconventional instrument and when they don't expect surprise and get a conventional instrument, and the slowest reading times when readers expect surprise but get a conventional instrument and when they don't expect surprise but get an unconventional instrument.

### 4.3.2 Methods

104 subjects were recruited via AMT for an SPR study on the webspr platform, hosted on Ibex Farm. Each subject was paid $0.60. Each subject saw only four critical stimuli, one from each of the four conditions; the number of subjects was chosen so that each individual stimulus was read by at least 8 subjects. An in-place display format was used so that subjects could not ascertain how close they were to the end of the sentence, forcing them to attend to linguistic cues about the distribution of information in the sentence.

Subjects reporting any native language other than English were excluded from analysis. Subjects with a mean RT four s.d. above the overall mean were excluded from analysis. Any trial with consecutive RTs under 50 ms, indicating that the subject was simply holding down a key, was also excluded from analysis. RTs above 3000 ms were excluded. Subjects scoring less 80% accuracy on comprehension questions were removed from the dataset. After this filtering, data from 84 unique subjects remained.

RT data was analyzed using ANOVAs and mixed-effects regression models with subject and instrument as random intercepts. For regression models, the dummy-coding took the *expect boring–conventional instrument* condition as the baseline, so all main effects and interactions should be interpreted as adding or subtracting from the average reading time in that condition.

### 4.3.3 Results

Figure 16 shows the average reading times at the critical instrument word and in the spillover region after it. The results from this experiment do not directly support the hypothesis that the violation of discourse expectations results in processing difficulty. The reasons for this failure will be covered in the discussion section below.

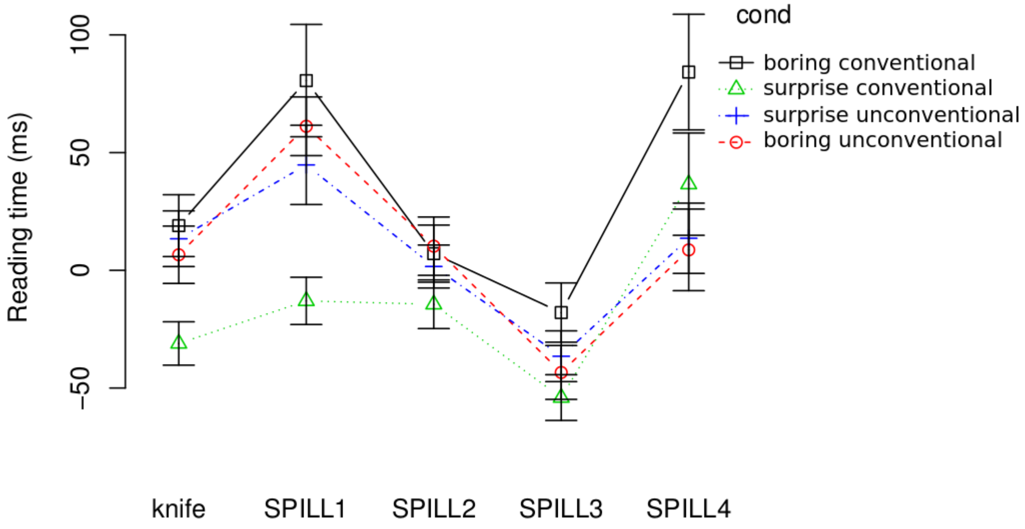

Figure 16: Mean reading times, residualized on word length, at the critical instrument word and following words. For conditions, *surprise* represents the "expect surprise" condition while *boring* represents the "expect boring" condition. *conventional* and *unconventional* refer to the conventionality of the instrument for the action in question.

At the critical instrument and afterwards, average reading time is fastest in the condition where readers expect surprise but don't get it. In fact, the distinction between that condition and all other conditions, for which reading times are higher, is the only significant effect in the results.

The RTs for all words at *knife* and subsequent spillover regions were collapsed. The condition in which discourse context led to an expectation of surprise (The prefix sentence *You'll never believe it!*) had overall faster reading times than the other discourse condition (*What a boring week.*), reflected in a main effect of discourse type ($F1(1{,}82) = 4.891$, $p = 0.03$; $F2(1{,}12) = 5.279$, $p = 0.04$). This main effect indicates that *all* material, regardless of its actual informativity, was read faster after this discourse context. When readers expected surprise, the surprising instrument was read more slowly, as indicated by an interaction between the expectation of surprise and the surprisingness of the instrument ($F1(1{,}79) = 5.882$, $p = 0.018$; $F2(1{,}12) = 5.829$, $p = 0.033$). Perhaps troublingly for the construction of our stimuli, no significant main effect for instrument surprisingness emerged: the unconventional instruments did not provoke an overall slower reading time than conventional instruments ($F1(1{,}82) = 2.472$, $p = 0.12$ n.s.; $F2(1{,}12) = 1.455$, $p = 0.251$ n.s.). Table 5 shows the details of a mixed-effects model predicting reading time at the critical instrument and spillover words from the discourse expectation of surprise and actual surprise, with random intercepts for subject and word.

| Coefficients: | Est. Std. | Error | $t$ value | Pr(>\|$t$\|) |
|---:|---|---|---|---|
| (Intercept) | 525.2 | 21.3 | 24.6 | <0.001 |
| unconventional instrument | -23.4 | 14.6 | -1.6 | 0.1 |
| expect surprise | -49.6 | 14.7 | -3.4 | <0.001 |
| unconventional instrument * expect surprise | 41.9 | 20.7 | 2.0 | 0.04 |

Table 5: Mixed-effects linear regression predicting the RT for the instrument word and subsequent words. The intercept represents the mean RT when readers do not expect surprise and do not get surprise. $p$ values obtained by Markov Chain Monte Carlo bootstrapping.

### 4.3.4 Discussion

The basic shape of the significant effects in the data is that, when readers expect informative discourse, they read unconventional instruments slowly and conventional instruments quickly; when readers do not expect informative discourse, they read everything slowly. Unconventional instruments, when preceded by surprising discourse context, were read faster than all instruments with boring discourse context, but this effect was not statistically significant. The results do not support our hypothesis that the expectation of informativity makes surprising material easier and unsurprising material more confusing.

We believe a number of factors due to flaws in the experimental design contributed to this counterintuitive result. In subsequent studies, we disentangle these factors and end up finding results that not only validate the original hypothesis about the effect of discourse context, but also clarify exactly what went wrong in this study.

The nature of the discourse expectations induced by our stimuli is somewhat unclear. After the set-up sentence *What a boring week*, we had assumed that readers would expect only boring, uninformative material. However, this discourse setup is itself anomalous–why would any speaker exclaim that a week was boring, then proceed to give examples? It might actually lend itself to being followed by informative material, perhaps as examples of the few interesting things that happened during the week. It is also not clear what the statement of boredom would entail. Overall, the discourse context is so vague that it is not clear what kind of expectations it produces. A more focused discourse context, which clarified which aspects of upcoming material would be uninformative, might address this issue. This is the subject matter of Experiment 3.

The use of an in-place display format also potentially obscured effects of the expectation of informativity. A reader may expect that *something* in the upcoming sentence will be informative; thus when she reads an uninformative word she simply assumes that the surprising material is coming later. The increased speed observed for the condition where discourse leads one to expect surprise (*You'll never believe it!*) may be a result of readers rushing through the text to find the interesting parts, slowing down after satisfactorily surprising information has been received (hence the interaction observed in the data). A violation of this expectation could only occur *after* the reader has moved beyond the region where she expected surprise, when she realizes that no surprising information will ever follow. If the reader expects surprise will occur anywhere in an the upcoming discourse, this effect can only emerge after the discourse is completely finished. Since readers using the in-place display format do not know if they are at the end of a discourse until the comprehension question appears, we have no opportunity to observe what happens when readers realize that *nothing* informative will ever appear in the remaining discourse. We do indeed find the expected slowdown at the end of the sentence in Experiment 3.

Another basic problem with this study was in the use of the cleft construction to direct readers' attention. The cleft construction itself is usually taken to contain a focus position, where discourse-new information is conveyed; for this reason it may itself induce an expectation of informativity. If this is the case, it would explain why conventional instruments provoked such a slow reading time in this position (in fact, the slowest of all reading times) even without the discourse expectation of informativity. If this is the case, then we should see that, in general, cleft constructions slow down reading times for conventional instruments. This will be demonstrated in Experiment 4.

Overall, the results of this experiment are consistent with the following explanation. The discourse context *You'll never believe it!* induces a global expectation of informativity, in that readers expect that something in upcoming discourse will be informative. This causes readers to read faster, in anticipation of upcoming informative material; once they receive it, they slow down. When they do not receive it, they continue reading in hopes of finding informative material down the line. In addition to this effect, the cleft construction creates a local expectation of informativity that, when readers have no global expectations, makes predictable instruments more anomalous.

## 4.4 Experiment 3

The second SPR experiment aimed to focus on the effects of discourse context by providing readers with clear information about where informative material might lie in discourse. The previous experiment induced only a vague expectation of informativity, without providing an opportunity to see what happens when speakers realize that the discourse does not conform to their expectations.

Since this experiment induces an expectation of informativity about upcoming sentences, we expect reduced reading times only when readers realize that there is no remaining discourse in which informative material might be imparted–that is, we only expect slowdowns for discourse violations at or near the end of each discourse. We expect a negative interaction between discourse expectations and surprising instruments, such that when readers expect surprise, they read surprising instruments more quickly and unsurprising instruments more slowly.

### 4.4.1 Materials

The same 13 action-instrument pairs from the previous experiment were used, this time with a clearer discourse context which alerted readers that the informative or uninformative material would have to do with instruments. Each of the 13 actions appeared in 4 conditions for a total of 52 stimuli. Example stimuli in the four conditions are:

**Expect surprise, get unconventional:** My friend John is a surprising person who never does things the way you'd expect. For instance, in order to brush his teeth, he was using a knife the day before yesterday.

**Expect surprise, but get conventional:** My friend John is a surprising person who never does things the way you'd expect. For instance, in order to chop some carrots, he was using a knife the day before yesterday.

**Expect boring, but get unconventional:** My friend John is a boring person who always does things the way you'd expect. For instance, in order to brush his teeth, he was using a knife the day before yesterday.

**Expect boring, get conventional:** My friend John is a boring person who always does things the way you'd expect. For instance, in order to chop some carrots, he was using a knife the day before yesterday.

In order to make these long stimuli less conspicuous, filler discourses were constructed providing a similar discourse structure. For instance one filler sentence is:

NYU is a trendy school where young actors and actresses often go to study. For example, Elizabeth Olsen studied there while filming movies on the side.

### 4.4.2 Methods

Materials were presented in a dashed display, so that readers were aware when they had reached the end of the discourse. 136 subjects were recruited over AMT; each subject was paid $0.60. Filtering procedures were the same as for Experiment 2. After filtering, data from 110 subjects remained.

### 4.4.3 Results

The results from this study confirm the hypothesis about the expectation of informativity. At the last word of each stimulus, when readers were predisposed to expect informative discourse, reading times were *slower* for sentences with predictable instruments than for sentences with unpredictable instruments. When readers did not expect surprise, sentences with unpredictable instruments yielded slower reading times. Mean reading times in the critical region are shown in Figure 17.

Reading times at the final word, but not elsewhere, show a significant negative interaction between discourse expectations of surprise and unconventional instrument ($F1(1{,}101) = 9.556$, $p = 0.003$; $F2(1{,}12) = 12.066$, $p = 0.005$). This indicates lower reading times when discourse expectations are in line with actual discourse, that is, when readers expect surprise and get surprise, as compared to the base case where readers expect boring material and get a conventional instrument. The effect size is large enough that the overall *fastest* reading times for the final word are when speakers have a fulfilled expectation of informativity. A mixed-effects model with subject and word as random intercepts, shown in Table 6, confirms the significant interaction.

The effects at other words follow the same numerical pattern as the effects at the final word, but these effects are not significant.

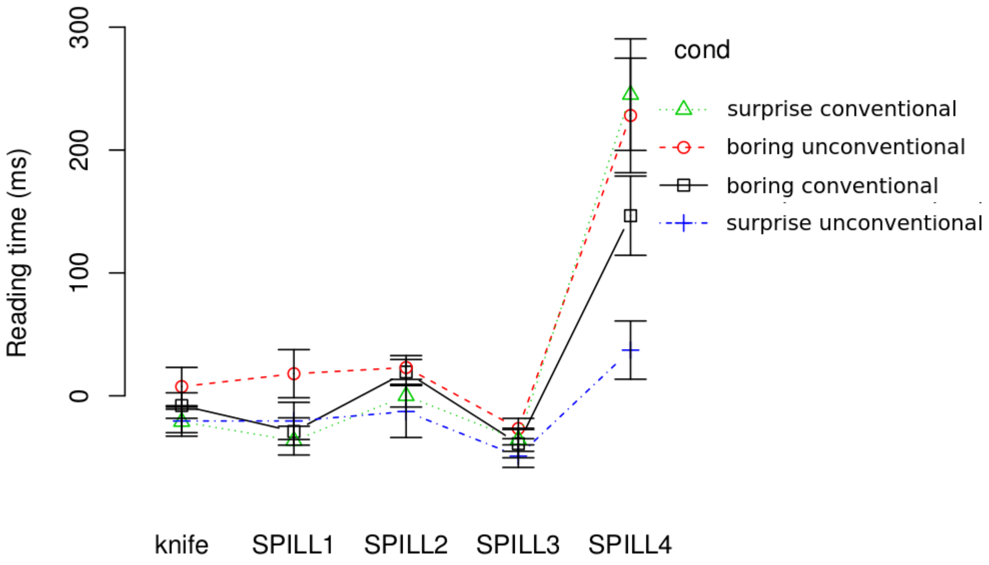

Figure 17: Mean reading times, residualized on word length, at the critical instrument word and following words. For conditions, *surprise* represents an expectation of surprise while *boring* represents no expectation of surprise. *Unconventional* indicates that the instrument word is actually surprising, while *conventional* indicates that it is not.

| Coefficients: | Est. Std. | Error | $t$ value | Pr(>\|$t$\|) |
|---:|---|---|---|---|
| (Intercept) | 686.4 | 41.1 | 16.7 | <0.001 |
| unconventional instrument | -20.0 | 17.5 | -1.1 | 0.1 |
| expect surprise | -35.8 | 17.7 | -2.0 | <0.001 |
| unconventional instrument * expect surprise | -82.8 | 17.7 | -4.7 | <0.001 |

Table 6: Mixed-effects linear regression predicting the RT for the final word of the stimulus. The intercept represents the mean RT when readers do not expect surprise and do not get surprise. $p$ values obtained by Markov Chain Monte Carlo bootstrapping.

### 4.4.4 Discussion

This experiment focused readers' expectations of informativity on instruments, and allowed them to know when the discourse was completed. At the end of the discourse, the condition with the lowest reading time was one in which readers expected surprise and got an unconventional instrument. All other conditions–those in which discourse expectations were violated, and those in which the discourse context was anomalous because it led readers not to expect informativity–were read more slowly.

Overall the results of this study support the hypothesized role of the expectation of informativity: it makes unpredictable material easier to process while making predictable material anomalous. The slow reading times in the condition where readers do not expect surprise might indicate that this is an anomalous kind of discourse to begin with.

The fact that these effects appeared significantly only at the final word could be for three reasons: (1) a lack of statistical power for detecting the effect at earlier words, (2) the effect of abnormal discourse only appears once readers know that nothing might follow that could potentially save the discourse coherence, or (3) the slowing effects of abnormal discourse appear only when readers are pondering the completed sentence, and the effect is not on on-line processing.

These findings are, to our knowledge, the first to find a situation where predictable material is *harder* to process than unpredictable material.They expand the effects of global discourse context, investigated in Nieuwland and van Berkum (2006), from changing expectations from one set of words to another to changing expectations about informativity. The results highlight the importance of taking discourse factors, such as the expectation of informativity, into account when theorizing about predictability effects in human language.

## 4.5 Experiment 4

The next study focuses on whether the same kind of effect as found in Experiment 3 can be induced by a conventional linguistic structure, rather than by discourse context. Specifically, we use wh-clefts to signal the location of informative material. Since the wh-cleft is a marked construction, hearers may assume that speakers use it for a reason, and if the material in the cleft is uninformative, such an utterance would be anomalous.

Since discourse context can lead readers to expect surprise anywhere in following material, a violation of that expectation only really occurs at the end of the discourse, when the reader knows nothing informative can follow. With a wh-cleft, the effect is localized, so that the effects of violated expectations should appear immediately within the cleft construction. We then expect to find the same general pattern of results as in Experiment 3, but localized at the instrument word. That is, we expect to find a negative interaction between the expectation of surprise and the actual surprisingness of an instrument, indicating that the expectation of surprise makes surprising material easier to process and unsurprising material harder to process.

### 4.5.1 Materials

The 52 stimuli from Experiments 2 and 3 were reworked so that a surprising instrument was presented either in a wh-cleft or not. For example:

**Cleft, get unconventional:** My friend John was digging a hole yesterday in the afternoon. What he was digging the hole with was a fork, and he almost stabbed himself.

**Cleft, but get conventional:** My friend John was eating some steak yesterday in the afternoon. What he was eating the steak with was a fork, and he almost stabbed himself.

**No cleft, but get unconventional:** My friend John was digging a hole yesterday in the afternoon. He was digging the hole with a fork, and he almost stabbed himself.

**No cleft, get conventional:** My friend John was eating some steak yesterday in the afternoon. He was eating the steak with a fork, and he almost stabbed himself.

The comma after the instrument provided the reader with a cue that the constituent inside the wh-cleft had terminated, and that the opportunity for informative material had passed. We thus expect the effects of violated discourse expectations to occur at the instrument noun itself, since this word is presented to readers as *fork,*.

### 4.5.2 Methods

Materials were presented in a dashed display. 136 subjects were recruited over AMT; each subject was paid $0.70. Filtering procedures were the same as for Experiments 2 and 3. After filtering, data from 99 subjects remained.

### 4.5.3 Results

For the critical region of the instrument itself, there were main effects of construction, whereby the cleft condition yielded slower reading times than the non-cleft ($p$ <0.001), and conventionality, whereby unconventional instruments yielded slower reading times than conventional instruments ($p$ <0.025). There was also the predicted construction x conventionality interaction whereby the slowdown associated with unconventional instruments was eliminated in the cleft condition ($p$ = 0.006). Mean RTs are shown in Figure 18.

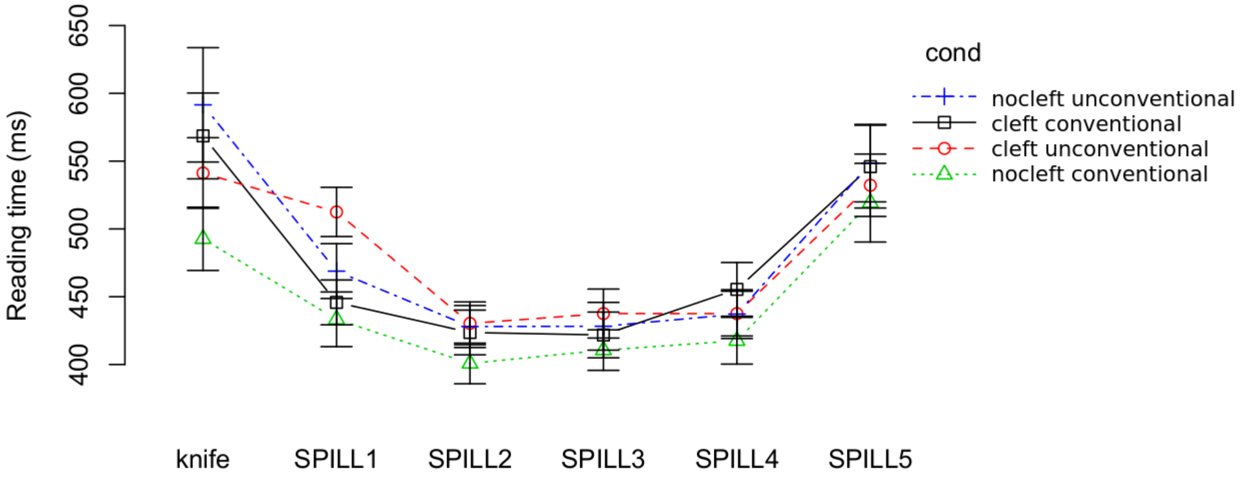

Figure 18: Mean reading times at the critical instrument word and following words. For conditions, *cleft* represents the presence of a wh-cleft while *nocleft* represents an instrumental PP. *unexpected* indicates that the instrument word is actually surprising, while *expected* indicates that it is not.

On the word immediately following the critical instrument word, the negative interaction disappears ($p$ = 0.4 n.s.), leaving only positive main effects of the cleft ($p$ <0.001) and of surprising instrument ($p$ <0.001). The high reading time visible at SPILL1 in Figure 18 for the unconventional instrument in the cleft construction is only an additive effect, where both unconventional instruments and clefts lead to lower RTs, with no significant interaction between these factors.

A mixed-effects model, with subject and word as random intercepts, summarizes these results at the instrument word in Table 7. Based on the coefficients from this model, it seems that an unconventional instrument causes a slowdown of 125 ms on average compared to a conventional instrument, while that same unconventional instrument in a wh-cleft causes a slowdown of only 69 ms. A conventional instrument in a wh-cleft, on the other hand, incurs a penalty of 77 ms.

| Coefficients: | Est. Std. | Error | $t$ value | Pr(>\|$t$\|) |
|---:|---|---|---|---|
| (Intercept) | 492.6 | 36.8 | 13.4 | <0.001 |
| unconventional instrument | 125.1 | 33.9 | 3.7 | <0.001 |
| cleft | 77.9 | 33.7 | 2.3 | 0.025 |
| unconventional instrument * cleft | -133.8 | 47.9 | 2.8 | 0.006 |

Table 7: Mixed-effects linear regression predicting the RT for the final word of the stimulus. The intercept represents the mean RT for a conventional instrument not in a cleft construction. $p$ values obtained by Markov Chain Monte Carlo bootstrapping.

### 4.5.4 Discussion

The same negative interaction as in Experiment 3 was observed in Experiment 4, at the predicted location of the instrument noun rather than at the final word of the discourse. The effect is weaker so there is no full crossover making predictable words harder to process than unpredictable ones, but the cleft construction does make unpredictable words easier to process and predictable words harder to process.

The results are in line with the hypothesis that speakers use an expectation of informativity in making predictions about upcoming words, and that the violation of this expectation can result in processing difficulty even for predictable words.

## 4.6 Summary and interpretation

The results overall indicate that an expectation of informativity can reduce the processing difficulty associated with unpredictable words, and that a violation of the expectation of informativity can make predictable words harder to process. The data are compatible with a number of different more detailed processing theories that take into account discourse expectations about informativity.

Experiment 1 validates the idea that discourse context can lead to greater informativity for instrument nouns, and that the expectation of informativity does not merely consist of shifting one's expectations from an uncertain set of conventional instruments to an equally uncertain set of unconventional instruments.

Experiment 2 fails to support its hypothesis, but it is possible to reinterpret the results from this study in light of subsequent findings. In that experiment, the fastest reading times were for conventional instruments in a discourse that yields expectations of informativity; all other reading times are slower than this one. The experiment likely failed because readers are unaware about *what* in the following discourse will be informative; when they encounter a conventional instrument, they simply continue reading at an accelerated pace in order to find the informative material. Since the experiment uses a centered display, there is no point where subjects know that the discourse is over and that their expectations have been violated, so there is no point where the effects of the violation of their expectations are evident in the data.

Experiment 3 finds the expected effect of discourse expectations at the final word, while Experiment 4 induces similar expectations with a conventional linguistic structure and finds the effects within that structure.

# 5 Conclusions and Future Directions

The experiments reported in this thesis provide evidence that speakers interpreting discourse have an expectation of informativity that shapes language processing. The main value in the thesis, in my mind, is to explore exactly how the expectation of informative discourse interacts with expectation-driven processing mechanisms. The results of the experiments show that when readers expect surprise, surprising material is easier to process and unsurprising material is harder to process. In some cases, the unsurprising material becomes even harder to process than the surprising material. To my knowledge these results are the first to show such an effect.

The closest results empirically to the present study are those in Corley et al. (2007), who studies surprisal by monitoring N400 effects after disfluencies. Since hearers expect words after disfluencies to be among that set of words that is hard to retrieve, a highly accessible word after a disfluency may case surprise–and Corley et al. (2007) indeed observe an N400 effect for frequent words after disfluencies. The present study finds a similar effect focusing on predictability, not frequency, with a correspondingly different interpretation. Hearers, noting that their interlocutor is speaking with a disfluency, may reason that the interlocutor is having processing difficulty and adjust their expectations accordingly. In our case, readers are noting that a writer is using certain discourse or conventional structures known to mark informative material, and they must reason about the writer's communicative intent, not about her level of processing difficulty. The current results are thus consistent with the results of Corley et al. (2007) and show that inverse predictability effects can appear even in language that is not marked by a disfluency, in the context of realistic discourse.

The results of this study join those of Nieuwland and van Berkum (2006) in showing that global discourse expectations can override local predictions, but they show it on a different level, in that readers' expectations *about informativity* are changed. Readers are not given a set of cues that makes an alternative, a priori unusual set of words likely; rather they must reason about what they would typically expect and then negate that expectation.

That discourse context can yield inverse predictability effects shows the importance of realistic discourse in psycholinguistic stimuli and models. Typical psycholinguistic stimuli are discursively anomalous declarative sentences, whereas most of the language that language users encounter is social discourse with the goal of communication. Greater attention should be paid to language as it is used in discourse, rather than 'language on holiday', used without communicative intent. The addition of only a sentence of discourse context in the studies of this thesis produced hitherto unexpected effects. Many more unexpected phenomena might appear with the psycholinguistic investigation of communicative discourse, and some psycholinguistic phenomena might turn out to be artifacts of the lack of apparent communicative intent in many psycholinguistic stimuli.

## 5.1 Implementational details

There are a few ways in which the expectation of informativity could shape human language processing in ways that are more or less compatible with the results of these experiments.

The expectation of informativity might only affect reading times by altering subjects' probability distributions over what words they believe will follow, in which case the standard effects of predictability (more predictable = faster RTs) would predict these findings. For instance, if a reader expects surprise, his probability distribution over following words might simply become flatter, such that the highly probable words become less so, while the less probable words become more so. The probability distribution would become closer to a uniform distribution. In that case, we would expect the expectation of informativity to make predictable words harder and unpredictable words easier, which is the finding in Experiment 4. However, this account would *not* predict the crossover effect where predictable words actually become harder than unpredictable ones. The crossover in Experiment 3 might contradict this prediction if it is interpreted as the result of on-line processing rather than post-hoc discourse integration. An implementation of the expectation of informativity through readers' probability distributions could only predict the crossover effect if speakers deliberately suppressed the most conventional continuations in their distributions.

Another implementation would involve the allocation of processing resources based on expectations about discourse informativity. A reader might allocate resources to make all processing easier, resulting in lower overall RTs for predictable words under an expectation of informativity. This account would seem to also predict a reduction in RT for unpredictable words in those positions, which contradicts results from Experiments 3 and 4. A resource-allocation account would only predict the observed rise in RT for predictable words if it allowed for processing difficulty due to misallocation of resources. A subject might, for instance, realize that resources were poorly allocated and then change them, with this reallocation process resulting in processing difficulty.

More generally, the results are compatible with theories of human language understanding involving the constant pragmatic calculation of the speaker's intentions. When a hearer notes that a speaker is using a cleft construction, she expects that the speaker has made the effort to use this marked construction for a reason; if the information in the cleft is uninformative, this calculation must be reassessed. Similarly, when a speaker speaks at all or uses an adjunct PP to mention an instrument at all, he must have a reason for this; an uninformative word, though predictable, may be pragmatically anomalous and lead to many difficult pragmatic calculations about the speaker's intentions. The results of these experiments show that these pragmatic considerations can interact with known predictability effects in interesting ways, in some cases overriding those effects.

## 5.2 Future directions

More immediately, the results of this thesis should be replicated in other paradigms and with other dependent variables. One obvious way that speakers signal the informational structure of their discourse is through prosody, which is not available in reading studies. Subjects might listen to sentences where words that are or are not informative are stressed; N400 responses might result from overly predictable words in such contexts. This would cement the claims of this thesis about processing difficulty, and clarify whether the effects are due to on-line or post-hoc processing. Targets other than instruments would also provide a valuable validation of the effect.

The paradigm of inducing expectations about informativity and then violating them also has the potential to cast light on the informational structure of language. Informativity in language is highly variable over time, and the studies of this thesis indicate that language comprehenders benefit from cues about where the most highly informative segments are to be expected. The pressures of information transfer likely lead languages to develop conventional structures to cue for informativity; since information transfer is a common goal of all languages, these cue structures might have universal characteristics. Existing typological tendencies and universals, such as given-before-new information ordering or the tendency of subjects to precede objects, might turn out to be a result of these pressures. The examination of cues to informativity in languages might also lead to the discovery of new typological generalizations.

# References

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

Arnold, J., Hudson Kam, C., and Tanenhaus, M. (2007). If you say thee uh- youre describing something hard: the on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33:914–930.

Arnold, J. E., Altmann, R., Fagnano, M., and Tanenhaus, M. K. (2004). The old and thee, uh, new. *Psychological Science*, pages 578–582.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Beattie, G. W. and Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, 22(3):201–211.

Bell, D. A. (1953). The internal information of english words. In Jackson, W., editor, *Communication Theory*, pages 383–391. New York: Academic.

Brown, P. and Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, 19:441–472.

Burton, N. G. and Licklider, J. C. R. (1955). Long-range constraints in the statistical structure of printed english. *American Journal of Psychology*, 68(4):650–653.

Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of EMNLP 2009.*

Chomsky, N. (1995). *The minimalist program.* MIT Press.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In Abner, N. and Bishop, J., editors, *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages 90–98, Somerville, MA, USA. Cascadilla Proceedings Project.

Cohen Priva, U. (Submitted 2011). Faithfulness as information utility. Stanford University manuscript.

Corley, M., MacGregor, L. J., and Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105:658–668.

Cover, T. M. and King, R. C. (1978). A convergent gambling estimate of the entropy of english. *IEEE Transactions on Information Theory*, IT-24(4):413–421.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory (2. ed.)*. Wiley.

Dagan, I., Lee, L., and Pereira, F. (1999). Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

Daltrozzo, J. and Schön, D. (2009). Conceptual processing in music as revealed by n400 effects on words and musical targets. *Journal of Cognitive Neuroscience*, 21(10):1882–1892.

Dayan, P. (2002). Matters temporal. *Trends in Cognitive Sciences*, 6(3):105–106.

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8:1117–1145.

Frank, A. and Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, pages 939–944, Washington, D.C.

Futrell, R. L. J. (2010). German noun class as a nominal protection device. Stanford University Senior Honors Thesis. Undergraduate Thesis.

Gahl, S. (2008). "thyme" and "time" are not homophones. the effect of lemma frequency on word durations in spontaneous speech. *Language*, 84(3):474–496.

Gahl, S. and Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 80:748–775.

Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*.

Genzel, D. and Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In Collins, M. and Steedman, M., editors, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, pages 65–72.

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

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, 10:96–106.

Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Language and Speech*, 4(4):232–237.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics 3: Speech Acts*, pages 43–58. New York: Academic.

Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309.

Hsueh, P. Y., Melville, P., and Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing.*

Ipeirotis, P. (2010). The new demographics of amazon mechanical turk.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001a). The effect of language model probability on pronunciation reduction. In *Proceedings of ICASSP-01 II, Salt Lake City, Utah*.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001b). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pages 229–254.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2002). The role of the lemma in form variation. In Warner, C. and Warner, N., editors, *Papers in Laboratory Phonology VII*, pages 1–34. New York: Mouton de Gruyter.

Kamide, Y., Altmann, G., and Haywood, S. (2003a). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *Journal of Memory and Language*, 49:133–159.

Kamide, Y., Scheepers, C., and Altmann, G. (2003b). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from german and english. *Journal of Psycholinguistic Research*, 32(1):37–55.

Keller, F., Gunasekharan, S., Mayo, N., and Corley, M. (2009). Timing accuracy of web experiments: A case study using the webexp software package. *Behavior Research Methods*, 41(1):1– 12.

Kidd, C., White, K. S., and Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, 14(4):925–934.

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5):580–602.

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

Kuperman, V. and Bresnan, J. W. (in press). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*.

Kutas, M. and Federmeier, K. D. (2009). N400. *Scholarpedia*, 4(10):7790.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307:161–163.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Levy, R. and Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *NIPS*, pages 849–856. MIT Press.

Mandelbrot, B. B. (1953). An information theory of the statistical structure of language. In Jackson, W., editor, *Communication Theory*, pages 503–512. New York: Academic.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, 189:226–228.

Marslen-Wilson, W. D. and Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8:1–71.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings NAACL-2010: Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA*.

Nieuwland, M. S. and van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, 358:1239–1253.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526.

Qian, T. and Jaeger, T. F. (submitted). Entropy profiles in language: A cross-linguistic investigation. *TBA*.

Ramscar, M., Dye, M., Popick, H. M., and O'Donnell-McCarthy, F. (2011a). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS ONE*, 6(7):e22501.

Ramscar, M., Suh, E., and Dye, M. (2011b). A steep price to pay? on the costs and benefits of learning relative pitch. In *Proceedings of the 33rd Meeting of the Cognitive Science Society*.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34:909–957.

Rayner, K., Binder, K. S., Ashby, J., and Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. *Vision Research*, 41:943–954.

Rayner, K. and Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14:191–201. 10.3758/BF03197692.

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

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3):151–160.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II: current research and theory*, pages 64–99. New York: Appleton-Century-Crofts.

Saussure, F. d. (1916). *Course in General Linguistics Ferdinand de Saussure*. Open Court Publishing Company.

Schnoebelen, T. and Kuperman, V. (2010). Using amazon mechanical turk for linguistic research. *Psihologija*, 43(4):441–464.

Schultz, W. (2007). Reward signals. *Scholarpedia*, 2(6):2184.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64.

Siegel, S. and Allan, L. G. (1996). The widespread influence of the rescorla-wagner model. *Psychonomic Bulletin & Review*, 3(3):314–321.

Sitnikova, T., Kuperberg, G., and Holcomb, P. J. (2003). Semantic integration in videos of real0world events: an electrophysiological investigation. *Psychophysiology*, 40:160–164.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2009). Cheap and fastbut is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Honolulu, Hawaii: Association for Computational Linguistics.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43(1):155–167.

Staub, A. and Charles Clifton, J. (2006). Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32:425–436.

Sturt, P., Pickering, M. J., and Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40:136–150.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., and Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, 1(2):147–165.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Harvard University Press.

Ward, G. and Birner, B. (2004). Information structure and non-canonical syntax. In Horn, R. and Ward, G., editors, *The Handbook of Pragmatics*, pages 153–174.

Wittgenstein, L. (1953). *Philosophical Investigations.* New York: Blackwell.

Yarlett, D. G. (2008). *Similarity-based generalization in language.* Dissertation, Stanford University.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* New York: Hafner.